UNIVERSIDAD NACIONAL DEL SUR

TESIS DE DOCTOR EN INGENIERÍA QUÍMICA

Metodologías Robustas de Reconciliación de Datos y
Tratamiento de Errores Sistemáticos

Claudia Elizabeth Llanos

BAHIA BLANCA ARGENTINA

2018

UNIVERSIDAD NACIONAL DEL SUR

TESIS DE DOCTOR EN INGENIERÍA QUÍMICA

Metodologías Robustas de Reconciliación de Datos y
Tratamiento de Errores Sistemáticos

Claudia Elizabeth Llanos

BAHIA BLANCA ARGENTINA

2018

PREFACIO

Esta Tesis se presenta como parte de los requisitos para optar al grado Académico de Doctor en Ingeniería Química, de la Universidad Nacional del Sur y no ha sido presentada previamente para la obtención de otro título en esta Universidad u otra. La misma contiene los resultados obtenidos en investigaciones llevadas a cabo en el ámbito de la Planta Piloto de Ingeniería Química, dependiente del Departamento de Ingeniería Química de la Universidad Nacional del Sur y del CONICET, durante el período comprendido entre el 10 de julio del 2013 y el 15 de febrero 2018, bajo la dirección de la Dra. Mabel Cristina Sánchez.

Claudia Elizabeth Llanos

UNIVERSIDAD NACIONAL DEL SUR
Secretaría General de Posgrado y Educación Continua

La presente tesis ha sido aprobada el …./…./….. , mereciendo la calificación de .....(……)

Certifico que fueron incluidos los cambios y correcciones sugeridas por los jurados.

Firma del Director

AGRADECIMIENTOS

Todas las personas que conocí en estos años han sido maravillosas, recibí mucho amor y respeto y espero poder devolverlo. Por eso doy gracias a Dios que nunca se equivoca y ha puesto las personas correctas en mi camino.

Quiero agradecer a mis padres que han cultivado en mí el amor por el trabajo; su dedicación y manos cansadas las llevo en mi corazón; gracias.

A mis hermanos y sobrina que me han regalado las mejores sonrisas y recuerdos de mi vida, los amo.

A mi padrino que cultivo en mi la curiosidad y ganas de disfrutar de un buen libro, de una buena historia.

A mis tías que me han acompañado en cada nuevo viaje o emprendimiento que he dado.

A mis abuelos que llenaron mi niñez de cuentos y caminatas, besos y caricias, y me enseñaron el valor de lo que no se ve pero se siente.

A mi amor que ha hecho que mis días sean más bonitos.

A mis compañeros de oficina con los que compartí 5 hermosos años y se han convertido en familia.

A mis amigos de PLAPIQUI con los que compartimos largas tardes de té, filosofía y folklore.

A mis compañeros del coro de la UNS y voluntariado que llenaron mi espíritu con sus locuras.

A mis amigos de Tucumán, de la Estación Experimental y Cátedra de Orgánica de la UNT que me impulsaron a seguir mis sueños.

A todas las personas que me acompañaron en este camino y en especial a mi directora de Tesis:

Gracias Mabel, por abrirme las puertas de tu casa y brindarte por completo en el desarrollo y escritura de ésta tesis.

A la educación pública de este país en la que tengo puestas todas mis esperanzas porque estoy convencida que es la herramienta que nos va a permitir construir un mejor futuro, una mejor sociedad.

¡Gracias Totales!

Todos tenemos derecho
a saber,

saber sirve para
participar

y hay que participar
para

construir un mundo
más justo

(L. Milani)

RESUMEN

La operación de las plantas químicas actuales se caracteriza por la necesidad de introducir cambios rápidos y de bajo costo con el fin de mejorar su rentabilidad, cumplir normas medioambientales y de seguridad, y obtener un producto final de una especificación dada. Con este propósito es esencial conocer el estado actual del proceso, el cual se infiere a partir de las mediciones y del modelo que lo representa.

A pesar de los recientes avances en la fabricación de instrumentos, las mediciones siempre presentan errores aleatorios y en ocasiones también contienen errores sistemáticos. El empleo de los valores de las mediciones sin tratamiento puede ocasionar un deterioro significativo en el funcionamiento de la planta, de allí la importancia de aplicar metodologías que conviertan los datos obtenidos por los sensores en información confiable.

La Reconciliación de Datos Clásica es una técnica probada que permite reducir los errores aleatorios de las mediciones. Con esta metodología se obtienen estimaciones más precisas de las observaciones, que son consistentes con el modelo. Sin embargo la presencia de errores sistemáticos invalida su base estadística, por lo que éstos deben ser detectados, identificados, y estimados o eliminados antes de aplicarla. Para evitar estos inconvenientes, se propusieron estrategias de Reconciliación de Datos Robusta (RDR) que son insensibles a una cantidad moderada de Errores Sistemáticos Esporádicos (ESE), dado que reemplazan la función Cuadrados Mínimos por un M-estimador.

En esta tesis se presentan nuevas metodologías de RDR que combinan las bondades de los M-estimadores monótonos y redescendientes. Se desarrolla un Método Simple que proporciona buenas estimaciones para las mediciones reconciliadas, y su carga

computacional es baja debido a que se lo inicializa con una mediana robusta de las observaciones.

Por otra parte, se formula el Test Robusto de las Mediciones (TRM) que utiliza la redundancia temporal provista por un conjunto de observaciones, y consigue detectar e identificar mediciones atípicas en variables con redundancia espacial nula, y con un porcentaje de aciertos idéntico al de las variables medidas redundantes. Esto es un notable avance en las técnicas de Detección de ESE pues independiza la capacidad de detección de la redundancia espacial. Además, el TMR permite identificar las variables con ESE en sistemas complejos, como procesos con corriente paralelas o variables equivalentes. En los mismos se logran aislar variables problemáticas sin generar falsas alarmas o perder capacidad de detección. Con lo cual se aborda un problema cuya solución estaba pendiente hasta el momento.

El efecto de la presencia de ESE puede ser contrarrestado por la RDR. No obstante, existen Errores Sistemáticos que Persisten en el Tiempo (ESPT), las estimaciones se ven deterioradas. En tal sentido, se presenta una nueva metodología para la detección y clasificación de ESPT basada en la técnica de Regresión Lineal Robusta y un procedimiento para el tratamiento integral de los errores sistemáticos que mejora significativamente la exactitud de las estimaciones de las variables.

Las estrategias propuestas en esta tesis han sido probadas satisfactoriamente en un proceso de mayor escala correspondiente a una planta de biodiésel. Se concluye que la correcta aplicación de la Estadística Robusta al procesamiento de datos permite desarrollar estrategias que proveen estimaciones insesgadas de las variables de proceso, con resultados reproducibles y aplicables a otros sistemas.

ABSTRACT

Nowadays, chemical plants need to introduce fast and low-cost changes in the operation to enhance their profitability, to satisfy environmental and safety regulations, and to obtain a final product of a certain quality. With this purpose, it is essential to know the current process state, which is estimated using the measurements and the model that represents its operation.

Despite the recent improvements in instruments manufacturing, measurements are always corrupted with random errors, and sometimes they also are contaminated with systematic ones. The use of untreated observations is detrimental for plant operation; therefore, it is important to apply strategies that transform the data given by sensors in reliable information.

The Classic Data Reconciliation (RDC) is a well-known technique that reduces the effect of random measurement errors. It provides more precise estimates of the observations, which are consistent with the process model. But the presence of systematic errors invalidates the statistical basis of that procedure. Therefore, those errors should be detected, identified, and estimated or eliminated before the application of RDC. To avoid this problem, Robust Data Reconciliation (RDR) strategies have been proposed, whose behavior is not affected by the presence of a moderate quantity of Sporadic Systematic Errors (ESE). They replace the Least Square Function by an M-estimator.

In this thesis, two RDR methodologies are presented which combine the advantages of monotone and redescendent M-estimators. The Simple Method is proposed, which provides unbiased estimates of the reconciled measurements. Its computation requirement is low because the procedure is initialized using a robust estimate of the observation median.

Furthermore, the Robust Measurement Test (TRM) is proposed. It uses the temporal redundancy provided by a set of measurements, and allows the detection and identification of atypical observations for measured variables which have null spatial-redundancy. Their identification percentages are similar to those obtained for the redundant measured ones. This a great advance in the ESE Detection area because for the new method the detection does not depend on the spatial-redundancy. Even more, TMR allows to identify ESE for complex systems, such as processes which have parallel streams and equivalent set of measurements. It isolates the measurements with ESE at a low rate of false alarms and high detection percentages. This has provided a solution to a subject unsolved until now.

Even though the detrimental effect of ESE can be reduced by the RDR, the presence of Systematic Errors that Persist in Time (ESPT) deteriorates variable estimates. In this sense, a new methodology is presented to detect the ESPT, and classify them using the Linear Robust Regression Technique. Also the treatment of all systematic errors is tackled using a comprehensive procedure that significantly enhances the accuracy of variable estimates.

The strategies proposed in this thesis have been satisfactorily proved for a plant of biodiesel production. It can be concluded that the right application of concepts from Robust Statistic to process data analysis allows to develop techniques which provide unbiased estimates, are reproducible and can be applied to other systems.

Índice General

Índice de Tablas

Índice de Figuras

Capítulo 1

Introducción

1 Introducción

1.1 Motivación

La operación de las plantas químicas actuales se caracteriza por la necesidad de introducir cambios rápidos y de bajo costo con el fin de mejorar su rentabilidad y cumplir normas medioambientales y de seguridad, al mismo tiempo que se obtiene un producto final de características específicas. Con este propósito se necesita conocer el estado actual del proceso, el cual está dado por los valores de las variables que lo conforman. La rápida y correcta interpretación de las mediciones permite conocer dicho estado actual y ejecutar, en consecuencia, acciones de control y optimización en línea, así como promover mejoras en el planeamiento gerencial de la planta.

A pesar de los recientes avances en la fabricación de instrumentos, las mediciones siempre presentan errores aleatorios y en ocasiones también contienen errores sistemáticos. Éstos son originados por el mal funcionamiento de los componentes del sistema de medición. El empleo de los valores de las mediciones sin tratamiento puede ocasionar un deterioro significativo en el funcionamiento de la planta, ya que afecta el correcto desempeño de los sistemas de control y puede llevar al proceso a operar en condiciones no económicas o inseguras. De allí la importancia de aplicar metodologías que conviertan los datos obtenidos por los sensores en información confiable sobre la operación actual del proceso, que pueda aplicarse posteriormente como entrada de los procedimientos de control, simulación y optimización en línea.

La Reconciliación de Datos Clásica (RDC) es una técnica utilizada para reducir los errores aleatorios de las mediciones. Esta metodología proporciona estimaciones más precisas de las observaciones, que son consistentes con las ecuaciones de modelo

empleado para representar la operación del sistema (Kuehn y Davidson, 1961). Al mismo tiempo, contribuye a reducir el número de análisis de laboratorio y la frecuencia de la calibración de los sensores (Lawrence, 1989).

Para que el procedimiento de RDC proporcione información útil del sistema es importante que los modelos matemáticos incorporen todo el conocimiento disponible del mismo. De manera general, se considera que el modelo está formado por variables relacionadas por medio de balances de masa, energía y cantidad de movimiento, además de algunas condiciones adicionales como restricciones físicas. Las variables se categorizan en medidas, **x**, o no medidas, **u**. Algunas variables medidas pueden ser corregidas por aplicación del procedimiento de RDC; a éstas se las denomina redundantes y las restantes son no redundantes. Además, las variables no medidas se dividen en observables y no observables. Solo las primeras pueden ser calculadas a partir de las ecuaciones del modelo y los valores corregidos de las observaciones (Romagnoli y Sánchez, 2000).

La RDC emplea la técnica de cuadrados mínimos (CM) con pesos y restricciones, las cuales comprenden el modelo que representa la operación del sistema. Si éste opera en estado estacionario y se considera que *N* es el número de mediciones disponibles de cada variable medida, se resuelve el siguiente problema de optimización:

$$[\hat{\mathbf{x}}_j, \hat{\mathbf{u}}_j] = \underset{x_j, u_j}{Min} \sum_{p=j-N+1}^{j} \sum_{i=1}^{I} \left(\frac{y_{ip} - x_{ij}}{\sigma_{y,i}} \right)^2$$

$$st.$$

$$\begin{aligned} \mathbf{f(x,u)} &= \mathbf{0} \\ \mathbf{h(x,u)} &\le \mathbf{0} \\ \mathbf{x^L} \le \mathbf{x} &\le \mathbf{x^U} \\ \mathbf{u^L} \le \mathbf{u} &\le \mathbf{u^U} \end{aligned} \qquad (1.1)$$

siendo x_{ij} el valor verdadero y desconocido de la *i*-ésima variable medida (*i*=1:*I*) en el *j*-ésimo intervalo de muestreo, y_{ip} es el valor de la observación de la *p*-ésima medición contenida en las *N* observaciones y $\sigma_{y,i}$ es el desvío estándar de la *i*-ésima medición que se asume conocido.

La solución óptima del problema anterior, $[\hat{\mathbf{x}}_j, \hat{\mathbf{u}}_j]$, comprende las estimaciones de las variables para el intervalo de muestreo *j*, obtenidas en función de las mediciones contenidas en una ventana de datos de tamaño *N* y del modelo del sistema. Éste se compone de los sistemas de ecuaciones de igualdad, **f**, de desigualdad, **h**, y de los límites para las variables medidas y no medidas. Se asume que las mediciones presentan sólo errores aleatorios y el modelo se conoce de manera exacta. Si alguna de estas suposiciones no se verifica, se pierde exactitud en los valores estimados de las variables (Narasimhan y Jordache, 2000).

Es práctica común asumir que los errores aleatorios, $\boldsymbol{\varepsilon}$, siguen una distribución normal multivariada con media igual al vector nulo y matriz de covarianza $\boldsymbol{\Sigma}$, es decir, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Dado que la presencia de mediciones atípicas infringe dicha suposición, se propusieron diferentes enfoques para reducir su efecto perjudicial sobre las estimaciones. Además, algunas de estas técnicas consideraron también la posibilidad de que el modelo del sistema no sea exacto debido a la existencia de pérdidas en las unidades del proceso.

Los test de hipótesis estadísticos son las técnicas más usadas para detectar y abordar situaciones que se apartan de las condiciones asumidas para aplicar la RDC. Algunos de los test estadísticos más empleados con este fin son el Test Global (Almasy y Sztano, 1975), el Test Nodal (Mah y co., 1976), el Test de las Mediciones (Mah y Tamhane, 1982), el Test de Máxima Potencia (Tamhane, 1982), el Test de Razón de Máxima

Verosimilitud (Narasimhan y Mah, 1987) y el Test de las Componentes Principales (Tong y Crowe, 1995). A diferencia del Test de las Mediciones o de Máxima Potencia, los restantes test necesitan de una técnica complementaria para identificar la posible causa de la anomalía. Por otra parte, los dos test previamente mencionados y el Test de las Componentes Principales necesitan que se haya realizado la RDC antes de ser calculados.

Por ejemplo, el Test Global en conjunto con la técnica que elimina una medición por vez permiten detectar la presencia de una medición atípica e identificar el sensor defectuoso. Éste se asocia a la observación cuya eliminación produce la máxima reducción del valor de la función objetivo del problema representado mediante la Ec. 1.1. El valor de dicha observación se estima en función de los valores reconciliados de las restantes utilizando el modelo del proceso. Por otra parte, la identificación de múltiples mediciones atípicas puede tratarse mediante estrategias seriales o simultáneas. La primera estrategia posibilita la rápida localización de los sensores defectuosos, pero lo hace con baja efectividad debido a los efectos de la dispersión de los errores entre las estimaciones de las variables (Keller y co., 1994; Kim y co. 1997). La segunda estrategia emplea un procedimiento combinatorio que insume mayor tiempo de cálculo (Sánchez, 1996). La estimación de los errores de las mediciones identificadas como atípicas forma parte del procedimiento de ambos tipos de estrategias, es decir, no se realiza después de la identificación. La posibilidad de estimar dichos errores depende de las características del sistema de ecuaciones redundantes. Éste se obtiene a partir del modelo del proceso y contiene sólo variables medidas redundantes.

De lo expuesto anteriormente resulta evidente que el uso del estimador CM requiere analizar si se satisfacen las suposiciones de la estrategia de RDC, y si esto no sucede, es necesario ejecutar otros procedimientos con el fin de obtener estimaciones confiables. En

tal caso, se demora la actualización de la información del estado del sistema, lo que podría impedir su aplicación en línea para procesos complejos.

A diferencia de la Estadística Clásica, la Estadística Robusta proporciona herramientas que aplicadas a la RDC permiten obtener estimaciones confiables de las variables del sistema aun cuando las mediciones siguen una distribución de probabilidad de manera aproximada (Maronna y co., 2006). En esta área se resaltan las contribuciones de Tuckey (1960, 1962) y Huber (1964, 1967) en la década del 60 y Hampel (1971, 1974) en la década del 70. La aplicación de estos avances teóricos fue posible gracias a la creciente disponibilidad y capacidad de los sistemas de cómputo.

En 1964 Huber presentó el concepto de M-estimadores, caracterizó una familia de estimadores más robusta que la función CM, y la utilizó para obtener estimaciones de localización, $\hat{x}$, cuando la distribución de las mediciones correspondía a una Normal Contaminada (Huber, 1964). Dichas estimaciones son la solución del siguiente problema de optimización:

$$\hat{x} = \arg\min_{x} \sum_{j=1}^{N} \rho(y_j - x) \tag{1.2}$$

donde arg min representa "el valor que minimiza", ρ la función del M-estimador, y N es el número de mediciones y_j disponibles de la variable medida x. Los M-estimadores son generalizaciones de la función de Máxima-Verosimilitud (Huber, 1981) que básicamente ponderan los residuos de las mediciones, otorgándole menor peso a aquellas que son atípicas.

La tesis de Hampel (1968) introdujo el concepto de la Función de Influencia, ψ, definida como la derivada de ρ. La ψ muestra la sensibilidad del estimador a los valores

de las observaciones. Su análisis permite clasificar a los M-estimadores en monótonos o redescendientes. Los primeros tienen una solución única, mientras que los segundos pueden tener más de una, por lo que necesitan de un buen punto de partida para alcanzar la solución correcta. Por otro lado, cuando se presentan errores de gran magnitud las funciones de pérdida con derivadas redescentientes son más robustas.

En el campo de la Ingeniería de Sistemas de Proceso, Tjoa y Biegler (1991) fueron los primeros autores que introdujeron conceptos de Estadística Robusta para la resolución del problema de reconciliación de datos. Su contribución consistió en reemplazar la función CM por la correspondiente a la Normal Contaminada en el problema de optimización representado por la Ec. 1.1. Esta formulación permitió considerar la presencia de mediciones con errores aleatorios y observaciones atípicas simultáneamente. Desde entonces, muchos otros autores discutieron y compararon la función CM con la antes mencionada, así como también con distintos M-estimadores (Alburquerque y Biegler, 1996; Arora y Biegler, 2001; Özyurt y Pike, 2004; Martinez Prata y co., 2008).

La Reconciliación de Datos Robusta (RDR) consiste en minimizar el valor de un M-estimador a la vez que se satisface el modelo que describe la operación del sistema, tal como representa la siguiente ecuación:

$$[\hat{\mathbf{x}}_j^R, \hat{\mathbf{u}}_j^R] = \underset{x_j, u_j}{Min} \quad \sum_{p=j-N+1}^{j} \sum_{i=1}^{I} \rho\left(\frac{y_{ip} - x_{ij}}{\sigma_{y,i}}\right)$$

$$st.$$

$$\begin{aligned} &\mathbf{f(x,u)=0} \\ &\mathbf{h(x,u)} \le \mathbf{0} \\ &\mathbf{x^L} \le \mathbf{x} \le \mathbf{x^U} \\ &\mathbf{u^L} \le \mathbf{u} \le \mathbf{u^U} \end{aligned} \tag{1.3}$$

La solución del problema de optimización, $[\hat{\mathbf{x}}_j^R, \hat{\mathbf{u}}_j^R]$, es una estimación confiable del estado del sistema aun cuando las observaciones de algunas variables contengan errores sistemáticos. Esto ocurre porque el M-estimador atenúa el efecto de las mediciones atípicas sobre las estimaciones, y en consecuencia, no es necesario detectar la presencia de errores sistemáticos como sucede con la RDC.

Los resultados de la RDR serán precisos siempre y cuando no se exceda el Punto de Quiebre (PQ) de la metodología. El PQ es la mayor proporción de valores atípicos que los datos pueden contener y aun así el estimador brinda información sobre el valor verdadero de la variable estimada (Huber, 1981). Es decir, el PQ es una medida global de reproducibilidad del estimador. Si para una muestra de tamaño *N*, el PQ es 1/*N*, esto es equivalente a decir que una sola observación puede distorsionar el estimador y hacer nula su utilidad práctica. Dicha condición se verifica para la función CM y ha motivado el desarrollo de estimadores con alto PQ.

Varios autores demostraron la efectividad de la RDR si las mediciones presentan Errores Sistemáticos Esporádicos (ESE). Se destacan los análisis realizados por Arora y Biegler (2001) y Özyurt y Pike (2004). En el primer trabajo se comparó la función CM con un M-estimador monótono, *Fair Function*, y otro redescendiente, Función de Hampel; este último fue inicializado con la solución del monótono y tuvo mejor comportamiento. El siguiente trabajo presentó un análisis de desempeño de siete funciones objetivo diferentes aplicadas al problema de RDR en estado estacionario y también incluyó tres criterios de detección de errores sistemáticos. Además estableció un parámetro que sirve como base para la comparación de todos los M-estimadores.

En la última década, los artículos publicados por Martinez Prata y co. (2008), Zhang y co. (2010) y Chen y co. (2013) emplearon los M-estimadores Welsch (WE), Cuadrados Mínimos Cuasi Pesados (CMCP) y Correntropía (CO), respectivamente, y analizaron el desempeño de dichos M-estimadores cuando las observaciones están contaminadas con ESE. En el trabajo de Chen y co. (2013) se mostró la superioridad del desempeño de la función CO con respecto al obtenido mediante las funciones CM y CMCP.

Asimismo, se simularon mediciones con Errores Sistemáticos que Persisten en el Tiempo (ESPT), cuya presencia puede ocasionar que se exceda el PQ de la metodología robusta. En tal sentido, Chen y co. (2013) simularon observaciones con sesgo y Nicholson y co. (2014) aplicaron RDR a un sistema dinámico cuyas mediciones presentaban sesgo y deriva. Ambos trabajos mostraron la efectividad de los estimadores en la corrección de las mediciones atípicas, pero no analizaron la detección y clasificación de los mismos, ni discutieron los problemas que surgen cuando se excede el PQ.

Por otro lado, Martinez Prata y co. (2010) presentaron una estrategia que utiliza el M-estimador de Welsch (WE) para la estimación robusta de las variables y su ψ para la detección de las mediciones atípicas. Estos autores distinguieron el error esporádico del sesgo analizando la persistencia del error en la variable. Asimismo, se diferenciaron mediciones con errores esporádicos, sesgo y deriva empleando la función CO. En tal sentido, Zhang y Chen (2015) presentaron una metodología para la detección de errores esporádicos basada en el cálculo de la distancia entre la última medición y las contenidas en la ventana de datos considerada. Además propusieron distinguir entre sesgo y deriva por medio del cálculo de la varianza de las mediciones y su comparación con una varianza umbral; aunque no explicaron cómo calcular este valor límite. Los dos trabajos descriptos

no realizaron un análisis integral del desempeño del método y tampoco analizaron el concepto del PQ.

En base a la discusión previa, se puede concluir que el procedimiento de RDR es útil en el ambiente industrial, ya que permite implementar la reconciliación de datos con mediciones crudas, sin necesidad de eliminar o corregir las observaciones con valores atípicos esporádicos. No obstante, se advierte que:

- Las metodologías robustas tienen un PQ, por lo que la persistencia de los errores sistemáticos puede ocasionar el deterioro de la estimación;

- No se han presentado estudios sistemáticos y exhaustivos que contemplen la presencia simultánea de mediciones con ESE y ESPT, ni metodologías reproducibles que permitan la detección e identificación de los mismos;

- No se han propuesto estrategias que utilicen de manera sinérgica las cualidades de los estimadores monótonos y redescendientes para lograr estimaciones exactas en tiempos cortos;

- No se han desarrollado estrategias reproducibles que simultáneamente detecten e identifique diferentes tipos de errores sistemáticos con alto desempeño, y que además permitan la corrección en línea de las mediciones con ESPT antes que se exceda el PQ.

En este contexto, la presente tesis tiene el propósito de avanzar en la resolución de los problemas citados

1.2 Objetivos

El objetivo general de esta tesis es el desarrollo de una estrategia basada en Estadística Robusta que proporcione estimaciones insesgadas de las variables, para su posterior aplicación en problemas de optimización en línea de plantas químicas, por lo tanto se considera que el proceso opera en estado estacionario.

Los objetivos particulares de este trabajo de investigación son:

- Desarrollar una técnica basada en RDR que permita disponer de estimaciones insesgadas de las variables del proceso con escaso tiempo de cómputo;
- Formular un test capaz de detectar e identificar simultáneamente las mediciones que presentan ESE con baja tasa de falsas alarmas y un elevado porcentaje de aciertos;
- Desarrollar una estrategia capaz de clasificar diferentes tipos de errores sistemáticos en los sensores (esporádico, sesgo o deriva) e implementar correcciones de las mediciones hasta tanto el sensor sea reparado;
- Comprobar el desempeño de la metodología propuesta en un bioproceso en el cual se disponga de mediciones de distintas magnitudes y observaciones con retardo.

1.3 Organización de la Tesis

Esta tesis está organizada de la manera descripta a continuación:

Capítulo 2: se realiza una revisión bibliográfica crítica sobre las metodologías existentes para abordar la detección e identificación de errores sistemáticos y resolver el problema de RDR;

Capítulo 3: se presentan dos metodologías robustas que permiten obtener estimaciones insesgadas de las variables de proceso empleando distintos tiempos de cómputo. Estas se comparan con las estrategias basadas en los M-estimadores cuyo uso se publicó exclusivamente durante la última década;

Capítulo 4: se presenta el Test Robusto de las Mediciones (TRM) y se analiza su desempeño en procesos cuya operación se representa mediante sistemas de ecuaciones algebraicas lineales y no lineales;

Capítulo 5: se desarrolla un algoritmo que hace uso de un método sencillo de RDR, desarrollado en el Capítulo 3, y del TRM para la reconciliación y detección de mediciones que presentan errores sistemáticos. A estas metodologías se le agrega una etapa de regresión robusta, la cual permite clasificar un ESPT en sesgo o deriva;

Capítulo 6. Se presentan los resultados de la aplicación del algoritmo desarrollado para una planta de biodiésel. El modelo de la misma es un sistema no lineal que contiene ecuaciones de conservación y variables medidas y no medidas;

Capítulo 7: Se presentan las conclusiones y sugerencias para trabajos futuros en esta temática.

1.4 Notación

$\mathbf{f}$	Sistema de restricciones de igualdad
$\mathbf{h}$	Sistema de restricciones de desigualdad
I	Número de variables medidas
N	Número de mediciones disponibles de la variable medida
$\mathbf{x}$	Vector de variables medidas
$\hat{\mathbf{x}}_j$	Vector reconciliado de las variables medidas en el intervalo *j*-ésimo
$\hat{\mathbf{x}}_j^R$	Vector reconciliado robusto de las variables medidas en el intervalo *j*-ésimo
$x_{i,p}$	Valor verdadero de la variable *i*-ésima en el intervalo *p*-ésimo
$\mathbf{x}^{\mathbf{U}}$	Límite superior las variables medidas
$\mathbf{x}^{\mathbf{L}}$	Límite inferior de las variables medidas
$\mathbf{u}$	Vector de variables no medidas
$\hat{\mathbf{u}}_j$	Vector estimado de las variables no medidas en el intervalo *j*-ésimo
$\hat{\mathbf{u}}_j$	Vector estimado robusto de las variables no medidas en el intervalo *j*-ésimo
$\mathbf{u}^{\mathbf{U}}$	Límite superior las variables no medidas
$\mathbf{u}^{\mathbf{L}}$	Límite inferior de las variables no medidas
$y_{i,p}$	Medición de la variable *i*-ésima en el intervalo *p*-ésimo
$\boldsymbol{\varepsilon}$	Vector de errores aleatorios
ρ	Función de pérdida del M-estimador
$\sigma_{y,i}$	Desvío estándar de la medición *i*-ésima
ψ	Función Influencia del M-estimador

$\boldsymbol{\Sigma}$	Matriz de covarianza de las mediciones
$\mathcal{N}$	Distribución normal

1.5 Acrónimos

CM	Cuadrados Mínimos
CMCP	Cuadrados Mínimos Cuasi Ponderados
CO	Correntropía
ESE	Errores Sistemáticos Esporádicos
ESPT	Errores Sistemáticos que Persisten en el Tiempo
RDC	Reconciliación de Datos Clásica
RDR	Reconciliación de Datos Robusta
PQ	Punto de Quiebre
WE	Welsch

Capítulo 2

Revisión Bibliográfica

2 Revisión Bibliográfica

2.1 Introducción

En este capítulo se presentan los resultados del análisis crítico de la bibliografía relacionada con el tratamiento de los errores sistemáticos, que invalidan las bases estadísticas de la Reconciliación de Datos Clásica (RDC), y la aplicación de Reconciliación de Datos Robusta (RDR) a las mediciones de procesos químicos. Se analizan las ventajas y limitaciones de las metodologías existentes. Esta recopilación sirve como punto de partida para los desarrollos presentados en este trabajo de tesis.

2.2 Revisión Crítica

2.2.1 Reconciliación de Datos Clásica

Las mediciones de un proceso siempre están sujetas a errores aleatorios, por lo tanto al introducirlas en las ecuaciones de conservación de masa y energía, éstas no se satisfacen en forma exacta. La RDC tiene como objetivo minimizar las discrepancias entre el modelo del proceso y dichas observaciones. Para esto se formula el problema de optimización presentado a continuación.

2.2.1.1. Formulación General del Problema de RDC

DEFINICIÓN 2.1: Se denomina Reconciliación de Datos al procedimiento mediante el cual se obtienen estimaciones precisas de las variables de un proceso, a partir de los valores de las mediciones, que sean consistentes con el modelo del mismo.

DEFINICIÓN 2.2: En ausencia de errores sistemáticos, el vector de mediciones del sistema en un dado tiempo, **y**, de dimensión *I* se define como:

$$\mathbf{y} = \mathbf{x} + \boldsymbol{\varepsilon} \tag{2.1}$$

siendo **x** el vector de valores verdaderos de las variables medidas y ε el vector de errores aleatorios de las mediciones. Normalmente se asume que:

- ε presenta distribución normal con media cero, es decir, $E(\boldsymbol{\varepsilon}) = \mathbf{0}$
- los sucesivos vectores de mediciones son independientes, esto implica que:

  $E(\boldsymbol{\varepsilon}_i\, \boldsymbol{\varepsilon}_j^T) = 0 \quad \forall\, i \neq j$

- La matriz de covarianza de los errores, $\boldsymbol{\Sigma} = E(\boldsymbol{\varepsilon}_i\, \boldsymbol{\varepsilon}_i^T)$ se asume conocida y definida positiva.

Se supone que $\varepsilon \sim \mathcal{N}(0, \Sigma)$ en base al Teorema Central del Límite (Draper y Smith, 1968). Se considera que un error está formado por componentes de diferentes fuentes, por lo tanto, su distribución de probabilidad tenderá a la distribución normal cuando el número de componentes aumenta, cualquiera sea la distribución de los mismos.

DEFINICIÓN 2.3: Las relaciones existentes entre las variables de un sistema que opera en estado estacionario pueden describirse por un modelo basado en principios de conservación de masa, energía y momento, así como también restricciones físicas de los materiales, condiciones de operación segura, etc. En general, estas restricciones se representan mediante sistemas de ecuaciones algebraicas de igualdad **f**, de desigualdad **h** y límites para las variables:

$$\begin{aligned} \mathbf{f}(\mathbf{x},\mathbf{u}) &= \mathbf{0} \\ \mathbf{h}(\mathbf{x},\mathbf{u}) &\le \mathbf{0} \\ \mathbf{x}^{\mathbf{L}} \le \mathbf{x} &\le \mathbf{x}^{\mathbf{U}} \\ \mathbf{u}^{\mathbf{L}} \le \mathbf{u} &\le \mathbf{u}^{\mathbf{U}} \end{aligned} \quad , \tag{2.2}$$

siendo **u** el vector de variables no medidas.

Considerando las definiciones y suposiciones anteriores, la RDC resuelve el problema de optimización definido por la Ec. 2.3. Su objetivo es minimizar la función Cuadrados Mínimos (CM), cuyo argumento es el ajuste estandarizado de las mediciones. La solución de este problema es un conjunto de mediciones corregidas ($\hat{\mathbf{x}}$) y estimaciones de las variables no medidas ($\hat{\mathbf{u}}$) que satisfacen estrictamente las restricciones del modelo.

$$\begin{aligned} [\hat{\mathbf{x}},\, \hat{\mathbf{u}}] = \underset{\mathbf{x},\mathbf{u}}{Min} & \left(\mathbf{y}-\mathbf{x}\right)^{T} \boldsymbol{\Sigma}^{-1} \left(\mathbf{y}-\mathbf{x}\right) \\ st. & \\ & \mathbf{f}(\mathbf{x},\mathbf{u}) = \mathbf{0} \\ & \mathbf{h}(\mathbf{x},\mathbf{u}) \le \mathbf{0} \\ & \mathbf{x}^{\mathbf{L}} \le \mathbf{x} \le \mathbf{x}^{\mathbf{U}} \\ & \mathbf{u}^{\mathbf{L}} \le \mathbf{u} \le \mathbf{u}^{\mathbf{U}} \end{aligned} \tag{2.3}$$

En procesos químicos, la aplicación del procedimiento de RDC se inicia con el trabajo de Kuehn y Davidson (1961), quienes efectuaron la corrección de las mediciones en sistemas lineales que relacionan sólo variables medidas. Dichos autores aplicaron el método de los multiplicadores de Lagrange (Edgar y Himmelblau, 1989) y obtuvieron una expresión analítica de los ajustes de las observaciones. Esta estrategia ha sido empleada con frecuencia en combinación con el método de Crowe y co. (1983), que permite la eliminación de las variables no medidas en sistemas lineales mediante el empleo de matrices de proyección.

Desde sus inicios, diversos autores han realizado aportes que permitieron extender la RDC a sistemas bilineales (Crowe, 1986; Sánchez y Romagnoli, 1996) y no lineales (Britt y Lueke, 1973; Pai y Fisher, 1988; Swartz, 1989). Dado que el problema de RDC en sistemas no lineales es en esencia un problema de optimización de Programación No Lineal, los avances producidos en el campo de estas técnicas se volcaron a la resolución de la RDC y estimación de parámetros en procesos químicos (Liebman y Edgar, 1988; Ramamurthi y Bequette, 1990).

Se destaca la revisión general de RDC publicada por Crowe (1996).Además, los libros de Mah (1990), Madron (1992), Romagnoli y Ránchez (2000), y Narasinham y Jordache (2000) presentan una extensa descripción de las metodologías desarrolladas en esta área del conocimiento.

Los procedimientos de RDC se utilizaron en diversas plantas químicas. Por ejemplo, Van Der Heijden y co. (1993a y 1993b) aplicaron esta técnica en un proceso de fermentación, obteniendo como resultado práctico el aumento de conversión. Singh y co. (2001) lo emplearon para monitorear una planta de procesamiento de mineral. Sunde y Berg (2003) abordaron el problema para el circuito de aguas de una planta nuclear. Por su parte, Matyus y co. (2003) y Jacob y Paris (2003) resolvieron la RDC para un conjunto de datos reales de flujos de residuos urbanos y de una planta de pulpa y papel, respectivamente. Más recientemente, Martinez Prata y co. (2010) aplicaron RDC a un reactor de polimerización industrial, mientras que Rafiee y Behrouzshad (2016) lo hicieron en una planta de procesamiento de gas natural. Estos son algunos de los diversos trabajos que destacan la viabilidad del procedimiento de RDC para un amplio espectro de aplicaciones.

La RDC asume que el modelo de las observaciones representado mediante la Ec. 2.1 se satisface exactamente. De allí la importancia del valor de la medición, puesto que de todos los posibles conjuntos $\hat{\mathbf{x}}$ que son consistentes con el modelo del sistema, se selecciona el que produce la corrección más pequeña posible de las observaciones, teniendo en cuenta las propiedades estocásticas de las mediciones, que se incorporan mediante su matriz de covarianza $\boldsymbol{\Sigma}$.

Sin embargo, existen mediciones atípicas; estas son observaciones que no se ajustan al modelo representado por la Ec. 2.1. Además del error aleatorio estas mediciones presentan errores sistemáticos. Los más frecuentes son los Errores Sistemáticos Esporádicos (ESE), los sesgos y las derivas. A diferencia de los esporádicos, los dos últimos son Errores Sistemáticos que Persisten en el Tiempo (ESPT). A continuación se presentan los modelos de las mediciones atípicas más frecuentes.

DEFINICION 2.4: La *i*-ésima observación del vector **y**, presenta un error sistemático esporádico, denominado "outlier" en inglés, si:

$$y_i = x_i + \varepsilon_i + K_i \sigma_{y,i} \tag{2.4}$$

La constante K_i es representativa de la magnitud y signo del outlier; el valor de este error es muchas veces mayor que el desvío estándar de la *i*-ésima observación, $\sigma_{y,i}$.

DEFINICION 2.5: La *i*-ésima observación del vector **y** presenta un ESPT denominado sesgo si:

$$y_i = x_i + \varepsilon_i + \mathrm{B}_i(t)\sigma_{y,i}\,, \tag{2.5}$$

donde $B_i(t)$ es la magnitud del sesgo. Ésta se mantiene constante en el tiempo, es decir, $B_i(t)=B_i$. El error de la medición tiene una distribución $\mathcal{N}(B_i, \sigma_{y,i})$

DEFINICION 2.6: La *i*-ésima observación del vector **y** presenta un ESPT denominado deriva si:

$$y_i = x_i + \varepsilon_i + m_{drift,i}(t)\sigma_{y,i}\,, \tag{2.6}$$

donde $m_{drift,i}$ (t) representa la funcionalidad del error con el tiempo.

Si bien existe un número pequeño de errores sistemáticos presentes en un conjunto de mediciones, su presencia invalida la base estadística del procedimiento de RDC. Por lo tanto, es necesario detectar la presencia de estos errores, identificar su ubicación y eliminarlos o corregirlos con el fin de obtener estimaciones insesgadas de las variables del sistema.

2.2.1.2 Tratamiento de Mediciones con Errores Sistemáticos

En general, las técnicas utilizadas para el tratamiento de mediciones atípicas se basan en la aplicación de test de hipótesis estadísticos. En estas pruebas se plantea una hipótesis nula H_0, referida a la ausencia de errores sistemáticos, y una alternativa H_1, que considera su posible presencia. Se formula un estadístico, τ , que sigue una distribución conocida cuando se satisface H_0, y permite declarar la presencia de un error sistemático si su valor excede el crítico τ_c. Éste se selecciona de la tabla de valores de la función de distribución acumulada de τ dado un nivel de significancia α .

El desempeño de un test de hipótesis no es perfecto. Un τ puede declarar la presencia de un error sistemático cuando no existe; en este caso se dice que el test comete un Error Tipo 1 (ET1) o da una falsa alarma. Por otro lado el test puede declarar que no hay error sistemático cuando el error existe; en ese caso se produce un Error Tipo 2 (ET2). La potencia de un test estadístico, que es la probabilidad de detectar correctamente un

error, es igual a [1-Probabilidad (ET2)]. Los ET1 y ET2 están relacionados, por lo que al momento de seleccionar un α se busca balancear la capacidad de detección con la de falsas alarmas (Canavos, 1988).

Las estrategias basadas en los test de hipótesis han utilizado la redundancia espacial (RE) presente en un conjunto de ecuaciones algebraicas lineales que comprende sólo variables medidas redundantes. Estas se obtienen luego de eliminar las variables no medidas y las mediciones no redundantes empleando, por ejemplo, la descomposición ortogonal QR (Romagnoli y Sánchez, 2000). La matriz resultante, **A**, es representativa del conjunto de ecuaciones de reconciliación. Sólo las variables relacionadas por esta matriz pueden ser corregidas por la RDC. El sistema lineal resultante se representa mediante la siguiente ecuación:

$$\mathbf{A}\mathbf{x}_\mathbf{r} - \mathbf{c} = 0 \tag{2.7}$$

siendo **c** un vector de constantes de dimensión *M*, **A** una matriz de dimensión compatible [*M* x *I*], y $\mathbf{x_r}$ el vector de variables medidas redundantes cuyas mediciones están contenidas en un subvector de **y**, que se denota $\mathbf{y_r}$. Por simplicidad, $\mathbf{x_r}$ e $\mathbf{y_r}$ se renombran como **x** e **y**, respectivamente.

2.2.1.3 Test Clásicos

A continuación, se describen los principales test utilizados en conjunto con estrategias para detectar e identificar mediciones con ESE en el marco de la RDC.

Almasy y Sztano (1975) y Mah y co. (1976) propusieron el Test Global y el Test Nodal, los cuales se basan en el cálculo del vector de residuos, **r**, y su matriz de covarianza, **V**, que se definen a continuación:

$$\mathbf{r} = \mathbf{A}\mathbf{y} - \mathbf{c} \tag{2.8}$$

$$\mathbf{V} = \mathrm{Cov}(\mathbf{r}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T \tag{2.9}$$

Test Global (TG): test multivariado que combina todos los elementos del vector **r** en un único estadístico $\boldsymbol{\gamma}$

H_0: No hay errores sistemáticos en el vector **y**

H_1: Hay errores sistemáticos en el vector **y**

$$\boldsymbol{\gamma} = \mathbf{r}^T\mathbf{V}^{-1}\mathbf{r} \tag{2.10}$$

Si se satisface H_0, $\boldsymbol{\gamma} \sim \chi^2_\nu$, siendo ν el número de grados de libertad, que es igual al rango de la matriz **A**.

H_0 se rechaza si: $\boldsymbol{\gamma} \geq \chi^2_{1-\alpha,\nu}$

Este test es colectivo, por lo que una vez que se ha detectado la presencia de una situación anómala, se requiere otro procedimiento para la identificación de la variable que origina el problema.

Test Nodal (TN): se formulan tantos estadísticos $\tau_{r,m}$ como ecuaciones contenga el sistema dado por la Ec. 2.7

H_0: No hay errores sistemáticos en las mediciones de las variables medidas contenidas en la restricción *m*

H_1: Hay errores sistemáticos en las mediciones de las variables medidas contenidas en la restricción *m*

$$\tau_{r,m} = \frac{|r_m|}{\sqrt{V_{mm}}} \quad m = 1, 2, \ldots M \tag{2.11}$$

Si se satisface H_0, $\tau_{r,m} \sim \mathcal{N}(0,1)$. H_0 se rechaza si: $\tau_{r,m} > \tau_{c,1-\beta/2}$ siendo β el nivel de significancia derivado de la desigualdad de Sidak (Mah and Tamhane, 1982).

$$\beta = 1-(1-\alpha)^{1/M} \tag{2.12}$$

Esta prueba presenta los mismos inconvenientes que el TG, es decir necesita de una estrategia adicional que permita identificar la variable medida con error sistemático, aunque esta búsqueda se realiza en un conjunto menor de variables. El cálculo de un estadístico para cada ecuación puede originar problemas de detección debido a la cancelación de errores de magnitudes opuestas en variables contenidas en una misma ecuación.

Años más tardes Mah y Tamhane (1982) presentaron el Test de las Mediciones, TM, basado en el ajuste de las mediciones, **a**, y su matriz de covarianza **Q**:

$$\mathbf{a} = \mathbf{y} - \hat{\mathbf{x}} = \mathbf{\Sigma A}^T \mathbf{V}^1 \mathbf{r} \tag{2.13}$$

$$\mathbf{Q} = \mathrm{Cov}(\mathbf{a}) = \mathbf{\Sigma A}^T \mathbf{V}^{-1} \mathbf{A \Sigma}^T \tag{2.14}$$

Test de las Mediciones (TM): se formulan tantos estadísticos $\tau_{a,i}$ como variables medidas redundantes

H_0: No hay errores sistemáticos en el vector **y**

H_1: Hay errores sistemáticos en el vector **y**

$$\tau_{a,i} = \frac{|a_i|}{\sqrt{Q_{ii}}} \quad i = 1,2...I \tag{2.15}$$

Si se satisface H_0, $\tau_{a,i} \sim \mathcal{N}(0,1)$. H_0 se rechaza si: $\tau_{a,i} > \tau_{c,1-\beta/2}$; β se obtiene reemplazando *M* por I en la Ec. 2.12

A diferencia de las dos pruebas anteriores, este test asocia a cada variable un valor de estadístico, por lo que no necesita técnicas de identificación posterior. No obstante, la evaluación del vector de ajuste requiere que previamente se efectúe el procedimiento de RDC a fin de obtener la estimación de $\hat{\mathbf{x}}$. Este resultado se ve afectado por la presencia de errores sistemáticos lo cual deteriora el desempeño del test provocando un aumento de falsas alarmas.

Mah y Tamhane (1982) derivaron el Test de Máxima Potencia de las Mediciones (TMP) para sistemas lineales. Este test es útil para matrices de covarianza no diagonales y es capaz de detectar una medición atípica si sólo hay un único ESE presente en **y**. El estadístico del test se formula como:

$$\tau_{d,i} = \frac{|d_i|}{\sqrt{Q_{d,ii}}} \tag{2.16}$$

siendo:

$$\mathbf{d} = \boldsymbol{\Sigma}^{-1}\mathbf{a} \tag{2.17}$$

$$\mathbf{Q}_d = Cov(\mathbf{d}) = \mathbf{A}^T(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^T)^{-1}\mathbf{A} \tag{2.18}$$

Los test indicados previamente se aplican solamente a la detección de errores sistemáticos presentes en las mediciones. No indican la presencia de otros errores, por ejemplo errores en el modelo del proceso tales como una pérdida, conocida por su nombre en inglés como “leak”. Éstas pueden ser modeladas de la siguiente forma:

DEFINICION 2.7: El *m*-ésimo equipo o conjunto de equipos presenta una pérdida si las ecuaciones de reconciliación, obtenidas a partir de los balances de masa total del sistema, son:

$$\mathbf{Ac_M} = L \times \boldsymbol{\delta}_m \tag{2.19}$$

donde $\mathbf{c_M}$ representa el vector de caudales másicos medidos y redundantes, L es la magnitud de la falla y $\boldsymbol{\delta}_m$ es un vector de ceros a excepción de la posición m-ésima en la cual contiene un 1.

Ésta deficiencia fue superada por el método de Relación de Máxima Verosimilitud, propuesto por Narasimhan y Mah (1987), que permite detectar cualquier error que pueda ser matemáticamente modelado. Éste se representa como un vector $\mathbf{e}_{pm}$.

Test de Relación de Máxima Verosimilitud (TRMV): se formulan estadísticos basado en la relación de verosimilitud

H_0: No hay errores sistemáticos en el vector **y** ni en el modelo del proceso

H_1: Hay errores sistemáticos en el vector **y** o en el modelo del proceso

$$\tau_{\mathrm{v},i} = \frac{\left(\mathbf{e}_{pm,i}^T \mathbf{V}^{-1}\mathbf{r}\right)^2}{\left(\mathbf{e}_{pm,i}^T \mathbf{V}^{-1}\mathbf{e}_{pm,i}\right)} \tag{2.20}$$

$$\tau_{\mathrm{v}} = \sup \tau_{\mathrm{v},i} \quad i = 1, 2, ..., k \tag{2.21}$$

donde $\mathbf{e}_{mp,i}$ es un vector que representa errores de las mediciones o del modelo.

Si H_0 se satisface, $\tau_{\mathrm{v}} \sim \chi_1^2$

H_0 se rechaza si: $\tau_v > \chi^2_{1,1-\beta}$; donde β se calcula con la Ec. 2.12, reemplazando M por el número de errores sistemáticos hipotéticos, k.

Como se mencionó con anterioridad, el TMP solo entrega resultados correctos cuando hay un único ESE presente en el vector medición. Para afrontar el escenario de múltiples errores esporádicos, Tong y Crowe (1995) aplicaron el Test de las Componentes

Principales. Este enfoque también resulta útil para los casos en los que la matriz de covarianza de los ajustes o residuos no es diagonal, es decir cuando los errores de las mediciones están correlacionados.

El Test de las Componentes Principales puede realizarse utilizando como base el vector de ajuste de las mediciones (CPTM) o el residuo de las restricciones. A continuación se presenta el desarrollo correspondiente al CPTM.

Con el objetivo de localizar la variable con errores sistemáticos se realiza una evaluación de la contribución de cada variable al estadístico del test. Para esto es necesario calcular la matriz de autovectores $\mathbf{U}$ y la de autovalores $\mathbf{\Lambda}$ obtenidas a partir de $\mathbf{V}$ $\left(\mathbf{\Lambda} = \mathbf{U}^T \mathbf{V} \mathbf{U}\right)$. Dado que $\mathbf{V}$ es singular, se consideran los g autovalores no nulos:

$$U_g = U(:,1:g) \tag{2.22}$$

$$\Lambda_g = \Lambda(1:g,1:g) \tag{2.23}$$

Test de las Componentes Principales (CPTM): se formulan estadísticos basados en la proyección del vector de ajustes en el espacio de las componentes principales (CP).

H_0: No hay errores sistemáticos en el vector **y**

H_1: Hay errores sistemáticos en el vector **y**

$$p_{a,i} = \left[(\mathbf{U}_g \mathbf{\Lambda}_g^{-0.5})^T \mathbf{a} \right]_i \quad i = 1,2,...g \tag{2.24}$$

Si se satisface H_0, $p_{a,i} \sim \mathcal{N}(0,1)$. H_0 se rechaza si: $p_{a,i} > \tau_{c,1-\beta/2}\,;\beta$ se obtiene reemplazando M por g en la Ec. 2.12

El test identifica las CP que superan el valor crítico, pero necesita una estrategia extra que le permita localizar la o las variables con error sistemático. Este último procedimiento puede introducir error (Narasimhan y Jordache, 2000).

La comparación entre los 4 estadísticos presentados permite observar las siguientes similitudes y diferencias entre ellos:

- El TM y el TMP detectan e identifican la medición con error sistemático sin metodologías extra;
- Si se asume la presencia de un único error sistemático en las mediciones, entonces el TRMV es idéntico al TMP (Crowe,1989a; Narasimhan, 1990). Por otro lado, si se considera que pueden presentarse fallas en las mediciones o en el modelo del proceso, el TRMV tiene un mejor desempeño. No obstante, este último hace la suposición de que la falla se conoce y puede modelarse, lo cual no es siempre cierto;
- El TG presenta alguna ventaja respecto del TRMV. Además de que el cálculo del TG es sencillo, no necesita disponer de ningún tipo de conocimiento de la anomalía. Sin embargo, la información provista por el TG es limitada, pues éste sólo indica si hay falla o no, pero no señala en qué variable se presenta o la estimación de la misma;
- EL CPTM al igual que el TG, a menudo, detectan fallas que otros test no indican (Narasimhan y Jordache, 2000). Esto se debe a que son test multivariados. No obstante necesitan de una estrategia extra que les permita aislar el origen de la falla.

En base a esta comparación se observa que no existe un test que posea el mejor desempeño global y la mayoría de ellos necesitan utilizar estrategias extras que les permitan identificar la variable con medición atípica.

2.2.1.4 Estrategias de Detección de Errores Sistemáticos

Hasta aquí se han mencionado los principales test utilizados para la detección de errores sistemáticos. Exceptuando el TM y el TMP, los restantes test no realizan la detección e identificación simultáneas de las observaciones atípicas, motivo por el cual se han desarrollado estrategias específicas para localizarlas.

En tal sentido, se ha empleado la eliminación serial (Ripps, 1965; Serth y Hennan, 1986; Rosenberg y co., 1987). Éste es un procedimiento iterativo que identifica un error a la vez usando algún test estadístico, y elimina la medición correspondiente hasta que no se detecten más errores sistemáticos. La metodología tiene la desventaja de afectar la redundancia del sistema, pues las variables eliminadas pasan a formar parte del conjunto de variables no medidas, lo cual deteriora la precisión de la estimación. A continuación se citan algunos de los aportes más relevantes sobre esta técnica.

- Romagnoli (1983) propuso una metodología de eliminación serial basada en una búsqueda combinada en el sistema de ecuaciones y variables. Utilizó el TG para cuantificar el efecto de la eliminación de un error en la RDC. Además, formuló expresiones para estimar la magnitud del ESE luego de la identificación del conjunto de mediciones sospechosas;
- Serth y Heenan (1986) desarrollaron siete estrategias basadas en el TN y TM. Destacándose el Test de las Mediciones Iterativo Modificado (TMIM). En éste se retira una medición sospechosa por vez (eliminación serial), se corrigen las variables medidas y se verifica la condición de no negatividad de las variables (caudales). Si ésta última se cumple, entonces se elimina definitivamente la variable y se repite el procedimiento hasta que el test no detecte más mediciones defectuosas.
- Rosenberg y co. (1987) propusieron dos estrategias que combinan la eliminación serial y el TM, y agregan restricciones a las variables. El desarrollo tuvo como objetivo disminuir la cantidad de falsas alarmas del TM. Esto se logró a expensas de un incremento en el tiempo de cómputo y una disminución en la capacidad de detección.

Sin embargo, la eliminación de las variables medidas puede causar la no observabilidad del sistema. Además Crowe (1988, 1989b) demostró que la eliminación

secuencial de la medición más sospechosa en cada paso no conduce necesariamente a los verdaderos errores sistemáticos. Con el fin de superar los problemas de la eliminación serial surgieron las metodologías que se describen a continuación.

Narasimhan y Mah (1987) propusieron una técnica de Compensación Serial que fue utilizada en conjunto con el TRMV. Dado un conjunto de posibles ESE, el procedimiento identifica un error sistemático por vez, estima la magnitud de dicho error y compensa la correspondiente medición o ecuación de balance. La estimación secuencial de los errores sistemáticos origina inconvenientes cuando la presencia de un error afecta a otros errores presentes. Se pueden dar situaciones en donde las estimaciones de las variables presentan errores más grandes que las observaciones originales debido a inexactitudes en las estimaciones de la magnitud de los errores. Otros autores que utilizaron compensación serial fueron Bagajewicz y Jiang (1998) empleando el TM en sistemas dinámicos.

Rollins y Davis (1992) presentaron una estrategia denominada Técnica de Estimación Insesgada, más conocida por sus siglas en inglés como UBET. En esta metodología se supone que hay tantos errores sistemáticos como los que pueden ser estimados teniendo en cuenta la RE provista por el modelo del proceso, y además se asumen los tipos de errores y las variables afectadas. Se analizan todas las combinaciones de errores posibles, y se considera que aquella que produce el menor valor de la función objetivo corresponde a las variables que efectivamente tienen error. En ese trabajo no se propusieron estrategias de identificación propias.

Por su parte, Sánchez (1996) desarrolló una estrategia de dos etapas para sistemas lineales con variables redundantes denominadas SEGE por sus siglas en inglés. La primera etapa emplea el TG para seleccionar un grupo de variables y ecuaciones sospechosas. Luego se formulan todas las posibles combinaciones de errores sistemáticos

en las variables y se estiman las magnitudes del error. Aquella combinación que produzca la mínima función objetivo es la correspondiente a las variables con error sistemático. Esta estrategia tiene como desventaja la utilización de un procedimiento iterativo para la identificación de las mediciones atípicas. Sánchez y co. (1999) extendieron el SEGE para sistemas con variables no medidas, en ese mismo trabajo se repasó el concepto de errores equivalentes introducido por Bagajewicz y Jiang (1998).

DEFINICION 2.8: Dos conjuntos de errores sistemáticos son equivalentes cuando tienen el mismo efecto sobre la reconciliación de datos.

Esto implica que eliminar uno u otro conjunto lleva a obtener el mismo valor de la función objetivo, por lo que los conjuntos de errores sistemáticos equivalentes son teóricamente indistinguibles. En otras palabras, cuando se identifica un conjunto de errores sistemáticos existe la misma posibilidad de que la verdadera ubicación de los errores esté en cualquiera de sus conjuntos equivalentes (Bagajewicz, 2010). Romagnoli y Sánchez (2000) explicaron que esta situación ocurre cuando:

- La función objetivo es la misma para distintas combinaciones de variables con error;
- Existen problemas estructurales.

El mismo problema de identificación se presenta en distintas estrategias tales como UBET, CPTM y TRMV (Romagnoli y Sánchez, 2000; Narasimhan y Jordache, 2000).

El enfoque basado en analizar las CP también se utilizó junto con los métodos tradicionales (Jiang y co., 1999; Amand y co., 2001; Wang y co., 2002). Se destaca el trabajo de Jiang y co. (1999), en el que se compararon el UBET modificado, el SEGE y el TM seguido de eliminación serial con técnicas equivalentes que utilizan el CPTM para formular el conjunto de variables sospechosas. Se mostró que la incorporación de las CP

no mejora significativamente el comportamiento de las técnicas y en algunos casos hasta empeoró su desempeño. Recientemente, Sagar y co. (2015) presentaron el Test Iterativo de las Componentes Principales (TIPC), el cual fue comparado con los desempeños del TRMV y Test Iterativo de las Mediciones desarrollado por Serth y Hennan (1986). El TIPC es capaz de detectar un conjunto de variables sospechosas en las que se encuentran contenidas las mediciones con error sistemático. Para identificar correctamente las variables con observaciones atípicas, se realizan 2^{nk}-1 iteraciones, donde *nk* es el número de variables sospechosas.

En otros trabajos se propusieron metodologías que combinan dos test para hacer frente a las debilidades de cada prueba por separado. Yang y co. (1995) formularon el TM-TN para aprovechar las ventajas de ambos test, y disminuir el tiempo de cómputo de la búsqueda combinatorial asociada al uso del TN y la estrategia de Eliminación Serial. Este método fue reformulado por Wang y co. (2004) y Mei y co. (2006), quienes utilizaron un procedimiento iterativo para evitar la pérdida de redundancia del sistema. Otros autores aplicaron esta combinación de test con el fin de reducir el conjunto de variables sospechosas y de esta manera disminuir la complejidad del problema de optimización mezcla entera lineal formulado para realizar la identificación y estimación de los errores (Sun y co., 2010). Recientemente Zhou y Fu. (2016) utilizaron esta combinación de test clásicos para aislar secuencialmente las mediciones que presentan errores sistemáticos, y les asignaron menores pesos en el procedimiento de estimación.

Otra combinación estudiada fue la del TRMV y TN. Esta técnica utiliza una estrategia de detección y compensación serial (Jiang y co., 2011). Las medidas de desempeño alcanzadas con la combinación de los test superan las obtenidas con los test

individuales, pero se necesitan procedimientos iterativos para identificar las mediciones atípicas.

Todas las técnicas de detección e identificación mencionadas pueden utilizarse en procesos descriptos por modelos no lineales previa linealización de sus modelos. Sin embargo, esta estrategia puede ocasionar problemas en sistemas altamente no lineales. Por tal motivo, algunos autores realizaron modificaciones sobre las metodologías clásicas TMIM (Kim y co., 1997) y TRMV (Renganathan y Narasimhan, 1999).

De lo expuesto anteriormente se observa que:

- Los test estadísticos explotan la RE pero no utilizan la redundancia temporal;
- Pocos test logran detectar e identificar simultáneamente las mediciones atípicas, entre los que se destaca el TM. Sin embargo, el efecto de dispersión del error en las estimaciones de las variables medidas perjudica su desempeño, lo que provoca aumentos considerables en las falsas alarmas;
- Las estrategias que suponen la presencia de un error (TRMV, UBET) realizan iteraciones inútiles, pues se sabe que los errores sistemáticos se dan con baja probabilidad;
- Las estrategias desarrolladas con el fin de identificar y estimar simultáneamente la magnitud del error son técnicas combinatorias que aumentan el tiempo de cómputo, en especial cuando los errores sistemáticos son múltiples, y por lo tanto no son adecuadas para aplicaciones en línea;
- Es recomendable desarrollar estrategias de detección que se adapten tanto a sistemas lineales como a no lineales.

Las conclusiones previas muestran la necesidad de desarrollar estrategias robustas capaces de proporcionar estimaciones insesgadas y precisas aun cuando los datos presenten errores sistemáticos.

2.2.2 Reconciliación de Datos Robusta

DEFINICIÓN 2.9: Un procedimiento estadístico se denomina robusto si no es sensible a pequeños apartamientos de las suposiciones en las cuales se basó.

La aplicación de conceptos de Estadística Robusta tiene por objetivo obtener estimaciones insesgadas cuando los supuestos de la Estadística Clásica no se cumplen exactamente (Huber, 1981; Maronna y co., 2006).

DEFINICIÓN 2.10: Se denomina Reconciliación de Datos Robusta (RDR) al procedimiento mediante el cual se obtienen estimaciones insesgadas de las variables del proceso, que son consistente con su modelo, tanto cuando las mediciones siguen fielmente una distribución de probabilidad como cuando lo hacen de forma aproximada.

Las estrategias de RDR sustituyen el tradicional estimador de CM ponderados por una función objetivo que tiene en cuenta las contribuciones de las mediciones atípicas. Los primeros en abordar este problema fueron Tjoa y Biegler (1991), quienes emplearon inicialmente una función objetivo basada en la distribución Normal Contaminada (NC) siguiendo el principio de Máxima Verosimilitud, que se presenta a continuación:

Definición 2.11: La estimación de Máxima Verosimilitud de $\hat{x}_i$ se obtiene al maximizar la Función de Verosimilitud, $\mathcal{L}$, definida como la función de densidad de probabilidad conjunta de las mediciones:

$$\mathcal{L}(y_{i1}...y_{iN}, x_i) = \prod_{p=1}^{N} f_0(y_{ip} - x_i) \tag{2.25}$$

es decir,

$$\hat{x}_i = \arg\max_{x_i} \mathcal{L}(y_{i1}...y_{iN}, x_i) \tag{2.26}$$

donde arg max significa "el valor que maximiza".

Si f_0 se conoce exactamente, la estimación de Máxima Verosimilitud se considera óptima, ya que tiene asociada la menor varianza asintótica entre los estimadores no sesgados de x_i.

En el trabajo de Tjoa y Biegler (1991) se resolvió el siguiente problema de optimización:

$$\begin{aligned} [\hat{x}^R, \hat{u}^R] = \underset{x,u}{Min} \quad & \left\{ -\sum_{i=1}^{I} \ln\left[(1-p) e^{0.5 a_i^2 \sigma_{y,i}^{-2}} + \frac{p}{K_i} e^{0.5 a_i^2 \sigma_{y,i}^{-2} K^{-2}} \right] \right\} \\ st. \\ & \mathbf{f(x,u)=0} \\ & \mathbf{h(x,u)} \le \mathbf{0} \\ & \mathbf{x^L} \le \mathbf{x} \le \mathbf{x^U} \\ & \mathbf{u^L} \le \mathbf{u} \le \mathbf{u^U} \end{aligned} \tag{2.27}$$

donde la probabilidad de mediciones con errores sistemáticos es p (p<0.5). Cuando el procedimiento converge, una observación se identifica como un valor atípico si su contribución a la probabilidad de la muestra es mayor que la correspondiente al error aleatorio. Con esta formulación del problema no se requiere ejecutar la identificación de los ESE antes de realizar la RDR. El desempeño de este método depende en gran medida

de una adecuada caracterización del error, dada por los valores de los parámetros *p* y K_i, que en la práctica son desconocidos. Además, las funciones objetivo son frecuentemente no convexas y complejas, por lo tanto, existen problemas de convergencia del algoritmo de optimización (Albuquerque y Biegler, 1996).

Johnston y Kramer (1995) exploraron la analogía existente entre la RDC, ejecutada en combinación con una metodología de detección y eliminación de ESE, y la Regresión Robusta. Aplicaron el M-estimador denominado Función de Lorentz (Huber, 1981), que resulta insensible a la presencia de errores sistemáticos, siempre y cuando éstos no superen el Punto de Quiebre (PQ) de la metodología. Dichos autores resolvieron el siguiente problema de optimización:

$$
\begin{aligned}
[\hat{x}^R,\, \hat{u}^R] = \underset{x\,,u}{Min} \quad & \sum_{i=1}^{I} \rho\left(\frac{x_i - y_i}{\sigma_{y,i}}\right) \\
st. \\
& \mathbf{f(x,u)=0} \\
& \mathbf{h(x,u)} \le \mathbf{0} \\
& \mathbf{x^L} \le \mathbf{x} \le \mathbf{x^U} \\
& \mathbf{u^L} \le \mathbf{u} \le \mathbf{u^U}
\end{aligned}
\tag{2.28}
$$

donde ρ representa el M-estimador. Con posterioridad, Albuquerque y Biegler (1996) emplearon un M-estimador convexo, la Función Fair (FF), que tiene la interesante propiedad de converger al óptimo global y limitar el efecto de los ESE. Como este estimador no permite inferir de manera directa cuáles son las mediciones con valores atípicos, se utilizaron técnicas basadas en Estadística Exploratoria con este propósito.

Por su parte, Arora y Biegler (2001) aplicaron el Estimador Redescendiente en Tres Partes (ERTP) propuesto por Hampel (1974). Esta función anula el efecto de las mediciones atípicas lo que le provee una robustez superior a la FF. Dadas las características del ERTP, se necesita un buen punto de partida para que el problema de

optimización converja. Por tal motivo, se lo inicializó con la solución obtenida empleando la FF con M-estimador. La detección e identificación de valores atípicos se realizó empleando un punto de corte explícito. El problema de reconciliación se formuló empleando los vectores de observación disponibles en un horizonte de tiempo. Estos se organizan en una matriz Y_{ob}, de dimensiones fijas [*I* x *N*], conocida como ventana móvil, que contiene los últimos *N* vectores de medición recibidos (Liebman y co., 1992), es decir:

$$\mathbf{Y}_{ob} = \left[\mathbf{y}_{j-N+1}, \mathbf{y}_{j-N+2}, \ldots, \mathbf{y}_{j} \right] \tag{2.29}$$

donde *j* es el índice correspondiente al intervalo de muestreo actual. El problema de RDR queda formulado de la siguiente manera:

$$\begin{aligned} [\hat{x}_j^R,\, \hat{u}_j^R] = \underset{x_j, u_j}{Min} \quad & \sum_{p=j-N+1}^{j} \sum_{i=1}^{I} \rho\left(\frac{y_{ip} - x_{ij}}{\sigma_{y,i}} \right) \\ st. \quad & \\ & \mathbf{f(x,u)=0} \\ & \mathbf{h(x,u)} \le \mathbf{0} \\ & \mathbf{x^L} \le \mathbf{x} \le \mathbf{x^U} \\ & \mathbf{u^L} \le \mathbf{u} \le \mathbf{u^U} \end{aligned} \tag{2.30}$$

Cabe mencionar que los parámetros del estimador redescendiente se sintonizaron para el conjunto específico de datos minimizando el Criterio de Información de Akaike (CIA) mediante una búsqueda iterativa de la Sección Dorada.

Definición 2.12: El CIA es un estimador insesgado de la distancia relativa de Kullback-Leibler, el cual representa la cantidad de información perdida cuando se usa un modelo **f'** para aproximar a un modelo **f**. El CIA para un modelo dado está dado por la función de máxima verosimilitud $(\mathcal{L})$ y el número de parámetros estimables κ.

$$CIA = 2\sum_{i=1}^{N}\left[-\log(\mathcal{L}(\mathbf{a}(i,\kappa),i,\kappa)) + 2\dim(\kappa)\right] \tag{2.31}$$

Los parámetros calculados con esta técnica se adaptan a un conjunto de datos determinado, por lo cual al utilizar otro conjunto de mediciones, el desempeño deja de ser óptimo, y los parámetros deben recalcularse. Esto es propio de las técnicas adaptativas; estas tienen mayores requerimientos de cómputo, lo que perjudica su aplicación en línea.

Wang y Romagnoli (2003) propusieron un estimador parcialmente adaptativo basado en la distribución T Generalizada y un estimador totalmente adaptativo basado en la estimación no paramétrica de la función de densidad de probabilidad. Ambos procedimientos mostraron mayor robustez y eficiencia en comparación con los enfoques basados en las funciones CM y NC a expensas del aumento de la carga computacional, lo cual restringe su aplicación al análisis fuera de línea.

Por otra parte, se destaca la contribución de Özyurt y Pike (2004), ya que estos autores presentaron un análisis de desempeño de siete funciones objetivo, que se habían utilizado previamente para resolver problemas de RDR, y tres criterios de detección de ESE. Se consideraron procesos operando en estado estacionario, tanto simulados como industriales, para los que se obtuvieron resultados prometedores utilizando la distribución de Cauchy y el ERTP. Con fines comparativos, todos los M-estimadores se sintonizaron de forma tal que la Eficiencia Asintótica (*Ef*) de la M-estimación sea la misma.

Definición 2.13: La Eficiencia Asintótica de una M-estimación $\hat{x}_i$ es:

$$Ef(\hat{x}_i) = \frac{\nu_o}{\nu} \tag{2.32}$$

donde v_0 es la varianza asintótica de la estimación de Máxima Verosimilitud obtenida considerando la función CM y v es la varianza asintótica de la estimación cuando se emplea el M-estimador. La *Ef* mide cuánto se acerca la M-estimación al valor óptimo.

El trabajo de Özyurt y Pike (2004) marcó un precedente, pues estableció una base común para comparar el comportamiento de los M-estimadores, evitando así procedimientos adaptativos iterativos. Además se definieron puntos de cortes basados en la Estadística Robusta, tales como la regla de corte X84 y los mínimos y máximos de las derivadas de los M-estimadores.

Con posterioridad Martinez Prata y co. (2008) abordaron la RDR para procesos no lineales que operan en estado dinámico. Presentaron un análisis comparativo de desempeño que involucró el M-estimador de Welsch (WE) y las mismas funciones objetivo estudiadas por Özyurt y Pike (2004). Se utilizó como medida de desempeño la reducción del error en la estimación, y se concluyó que las funciones de Lorentz y WE proporcionaban los mejores valores reconciliados para los casos de estudio analizados.

$$
\begin{aligned}
[\hat{x}_j^R,\, \hat{u}_j^R] = \underset{x_j,u_j}{Min} \quad & \sum_{p=j-N+1}^{N} \sum_{i=1}^{I} \rho\left(\frac{y_{i,p} - x_i}{\sigma_{y,i}}\right) \\
st. \quad & \\
& \mathcal{D}\left(\frac{\partial \mathbf{y}(t)}{\partial t}, \mathbf{y}(t)\right) = \mathbf{0} \\
& \mathbf{f}(\mathbf{x},\mathbf{u}) = \mathbf{0} \\
& \mathbf{h}(\mathbf{x},\mathbf{u}) \le \mathbf{0} \\
& \mathbf{x}^{\mathbf{L}} \le \mathbf{x} \le \mathbf{x}^{\mathbf{U}} \\
& \mathbf{u}^{\mathbf{L}} \le \mathbf{u} \le \mathbf{u}^{\mathbf{U}}
\end{aligned}
\tag{2.33}
$$

Martinez Prata y co. (2008) simularon ESPT tipo sesgo y observaron que la presencia de los mismos afecta a la RDR provocando la pérdida de exactitud del vector estimado. No obstante, los autores no precisaron el número de simulaciones realizadas como para comprobar la representatividad de los resultados y los errores modelados se generaron siguiendo un único modelo de medición con error, que corresponde al de la distribución NC.

El M-estimador Cuadrados Mínimos Cuasi Ponderado (CMCP) fue formulado por Zhang y co. (2010). Esta función al igual que la FF acota el efecto del error sistemático. Los autores emplearon el Criterio de Información de Akaike con el fin de ajustar el parámetro del M-estimador para cada aplicación específica. Se comparó su desempeño con respecto al comportamiento de la FF y el ERTP en un caso de estudio lineal extraído de Serth y Hennan (1986), denominado Red de Ingreso de Vapor (SMN). El punto de corte se estableció fijando la probabilidad de cometer ET1 cuando las mediciones se ajustan exactamente a la distribución Normal. Los autores señalaron que el estimador CMCP resultó más efectivo que los otros M-estimadores. No obstante, no se indicaron los parámetros utilizados en la función ERTP, la cual es una función con derivada redescendiente que teóricamente debe tener mejor desempeño que la CMCP. Asimismo, se mencionó el uso del TM para detectar ESE, pero no se proporcionó información sobre cómo se lo formuló.

Más tarde, Chen y co. (2013) propusieron utilizar la función Correntropía (CO) como M-estimador. Su parámetro se calculó minimizando el Criterio de Información de Akaike. El punto de corte también se fijó seleccionando la probabilidad de cometer un dado ET1 cuando se satisface la distribución ideal. Se realizó un estudio de comparación de desempeño para el SMN, utilizando mediciones contenidas en una ventana de datos, y

se concluyó que la función CO tenía un desempeño superior al M-estimador CMCP. El procedimiento de detección de ESE no se presentó. Si bien se reportaron medidas de desempeño, se observa que la detección es siempre inferior al 50%.

Martinez Prata y co (2010) formularon el problema de RDR usando el M-estimador de WE para ajustar las mediciones y estimar los parámetros de un reactor de polipropileno industrial. Usaron el enfoque de ventana móvil para el tratamiento de los datos, y un algoritmo de optimización de Enjambre de Partículas. La estrategia se aplicó para la identificación de ESE y sesgos. Se utilizó un valor de corte del ajuste de las observaciones estandarizadas para determinar la presencia de valores atípicos. Los sesgos se identificaron como una secuencia de ESE del mismo signo. El número de intervalos de tiempo de esa secuencia se fijó utilizando el nivel de significancia del 5% de la distribución binomial. Si bien se propuso una metodología para detectar e identificar las variables con sesgo, no se planteó ninguna técnica para el tratamiento de los mismos.

Por su parte, Zhang y Chen (2015) abordaron la identificación de ESE, sesgos y derivas empleando el M-estimador CO. Los autores propusieron calcular un estadístico para cada medición en función de su ajuste y declararon la presencia de un error sistemático si ese estadístico era mayor que un valor crítico. Indicaron que usaban el estadístico del TM, sin embargo, calcularon el cociente entre el ajuste robusto y la varianza de la medición, lo cual no corresponde a la definición del TM. Además, establecieron un criterio de distancia-tiempo para distinguir entre los ESE y ESPT y utilizaron un umbral de la varianza muestral de los ajustes de la observaciones para discriminar entre sesgos y derivas, pero no indicaron qué criterio emplearon para fijar ese umbral, por lo cual los resultados no son reproducibles.

En la Tabla 2.1 se recopilan los artículos en los que se emplearon M-estimadores como función objetivo del problema de reconciliación de datos y se reportan los tipos de errores simulados en cada trabajo.

Tabla 2.1 M-estimadores usados como función objetivo de la RDR

AUTOR	M-ESTIMADOR	ERROR	AÑO
Tjoa y Biegler	Normal Contaminada	ESE	1991
Johnston y Kramer	Normal Contaminada y Lorentz	ESE	1995
Albuquerque y Biegler	Normal Contaminada y Fair	ESE	1996
Chen	Fair y Lorentz	ESE	1998
Arora y Biegler	Fair Hampel	ESE	2001
Özyurt y Pike	Normal Contaminada, Cauchy, Fair, Hampel, Logística y Lorentz	ESE	2004
Zhou	Huber	ESE	2006
Schladt y Hu	Normal Contaminada	ESE	2007
Alhaj-dibo y co.	Normal Contaminada	ESE	2008
Martinez Prata y co. (a)*[1]	Normal Contaminada, Andrews, Biweight, Cauchy, Fair, Hampel, Huber, Logística, Lorentz, Talwar y Welsch.	ESE y ESPT	2008
Martinez Prata y co. (b)*[2]	Normal, Fair, Hampel y Welsch	ESE y ESPT	2010
Zhang *et al**.	Cuadrados Mínimos Cuasi Ponderados	ESPT*	2010
Chen *et al**	Correntropía y Biweigth	ESPT*	2013
Nicholson *et al**[1].	Hampel	ESE y ESPT	2014
Chen *et al.**[2]	Correntropía	ESE y ESPT	2015

* Consideraron la presencia de sesgos de longitud igual a la ventana de datos;

*[1] Consideraron la presencia de ESPT pero no realizaron la detección de los mismos;

*[2] Realizaron la detección y clasificación de errores, pero no proporcionaron un análisis exhaustivo de su desempeño. Además las variables con ESPT fueron elegidas arbitrariamente.

En base a la búsqueda bibliográfica realizada, se ha comprobado que la RDR permite disponer de estimaciones insesgadas de las variables contenidas en un modelo cuando las mediciones están contaminadas con errores sistemáticos. No obstante, la persistencia de éstos en el tiempo provoca que el PQ de la metodología se exceda, lo cual disminuye la calidad de las estimaciones. Si bien el trabajo de Zhang y Chen (2015) trata la detección e identificación de los ESPT, la metodología no tiene una base estadística correcta.

En resumen, se observa que:

- El uso de estrategias adaptativas, por ejemplo la minimización del CIA empleando la búsqueda de la Sección Dorada, aumentan el tiempo de cómputo, por lo cual no resultan adecuadas para su empleo en línea;
- La detección de ESPT es fundamental para evitar la pérdida de exactitud en las estimaciones;
- El desempeño de las metodologías de RDR no se ha analizado para distintos modelos de errores sistemáticos;
- No se ha hecho una extensión correcta de los test basados en Estadística Clásica para considerar la presencia de ESE;
- No se han realizados análisis exhaustivos de la capacidad de detección de ESPT;
- No se han propuesto estrategias que permitan corregir las mediciones con ESPT una vez que los mismos se han identificado.

2.3 Conclusiones

La revisión bibliográfica ha evidenciado los avances realizado en el área de reconciliación de datos, así como también las temáticas pendientes de resolución. Cabe destacar que:

- La RDC se puede aplicar con éxito en procesos industriales. Se han desarrollado numerosos trabajos en los que se concluye que la RDC es una herramienta que mejora la precisión de las variables estimadas del proceso;
- El TM permite la detección e identificación simultáneas de ESE;
- Existen metodologías capaces de detectar y calcular la magnitud del error (TRMV, UBET, SEGE);
- Los test estadísticos pueden ser utilizados en sistemas no lineales;
- Las estrategias robustas son de interés dado que proporcionan estimaciones insesgadas aun cuando los errores de las mediciones no se distribuyen siguiendo exactamente una distribución normal;
- El uso de una ventana de datos mejora la precisión de las estimaciones;
- Si se fija la *Ef* de la M-estimación en ausencia de errores sistemáticos, es posible realizar comparaciones válidas del desempeño de diferentes M-estimadores;
- La RDR no necesita de métodos que compensen o eliminen las mediciones con error, siempre y cuando no se supere el PQ de la metodología.

Sin embargo se notan las siguientes falencias:

- Las publicaciones recientes en el área de RDC muestran la aplicación de la técnica a diversos sistemas industriales, pero no realizan aportes teóricos;

- La mayoría de las técnicas de detección de errores sistemáticos se basan en los primeros test desarrollados, que datan de la década del 80, los cuales utilizan estrategias iterativas para identificar las mediciones atípicas;
- El TM se ve perjudicado por la presencia de múltiples errores sistemáticos, lo cual aumenta la cantidad de falsas alarmas haciendo infactible su aplicación en procesos industriales;
- Las metodologías de detección que parten de la premisa de suponer la existencia de errores sistemáticos en el conjunto de mediciones (UBET, TRMV), en la mayoría de los casos, realizan cálculos para detectar y estimar errores inexistentes, pues se sabe que éstos se dan con baja frecuencia;
- La no linealidad de los modelos puede afectar el desempeño de los test. Por esto resulta conveniente desarrollar metodologías con elevado desempeño independientemente de la naturaleza del modelo;
- Las técnicas adaptativas requieren recursos de cómputo adicionales para estimar de manera robusta las variables del proceso;
- Se han desarrollados reglas de rechazo basadas en Estadística Robusta, que si bien no han tenido buen desempeño, constituyen los primeros intentos de un abordaje robusto del problema;
- Las metodologías robustas tienen un PQ que pude ser excedido si los errores sistemáticos persisten en el tiempo.

De las consideraciones previas surge la necesidad del desarrollo de estrategias, reproducibles y aplicables a todos los sistemas, capaces de proveer estimaciones insesgadas de las variables de una planta química que sirvan de entrada a los algoritmos de optimización en línea.

2.4 Notación

A	Vector de ajuste de las mediciones
A	Matriz representativa de las ecuaciones lineales de reconciliación
B	Magnitud del sesgo
C	Vector de constantes no nulas
D	Vector de ajuste estandarizado
e_{pm}	Error de la medición o del modelo
F	Sistema de restricciones de igualdad
f_0	Función de densidad de probabilidad
G	Número de autovalores no nulos
H	Sistema de restricciones de desigualdad
I	Número de variables medidas
K	Magnitud del ESE
L	Magnitud de la pérdida en un equipo
M	Número de ecuaciones del modelo
m_{drift}	Magnitud de la deriva
N	Número de réplicas de la variable medida
$\mathbf{p_a}$	Estadístico del CPTM
Q	Matriz de covarianza del ajuste
Q_d	Matriz de covarianza del ajuste estandarizado
r	Residuo
u	Vector de variables no medidas
$\hat{\mathbf{u}}$	Vector estimado de las variables no medidas
$\mathbf{u^U}$	Límite superior de las variables no medidas

$\mathbf{u^L}$	Límite inferior las variables no medidas
$\mathbf{U_g}$	Matriz de autovectores
$\mathbf{V}$	Matriz de covarianza del residuo
$\mathbf{y}$	Vector de mediciones
$\mathbf{x}$	Vector de variables medidas
$\hat{\mathbf{x}}$	Vector reconciliado de las variables medidas
$\mathbf{x^U}$	Límite superior de las variables medidas
$\mathbf{x^L}$	Límite inferior de las variables medidas
$\mathbf{x_r}$	Vector de variables medidas redundantes
$\mathbf{Y_{ob}}$	Matriz de observaciones
α	Nivel de significancia del test
β	Nivel de significancia dado por la desigualdad de Sidak
$\boldsymbol{\gamma}$	Estadístico del Test Global
$\boldsymbol{\delta}$	Vector de pérdidas
$\boldsymbol{\varepsilon}$	Vector de errores aleatorios
κ	Parámetros desconocidos
ν	Varianza de la M-estimación
ν_0	Varianza de la estimación obtenida con la función CM
$\sigma_{y,i}$	Desvio estándar de la medición i-ésima
τ	Estadístico
χ^2	Distribución chi cuadrado
$\boldsymbol{\Sigma}$	Matriz de covarianza de las mediciones
$\boldsymbol{\Lambda}_g$	Matriz de autovalores no nulos

$\mathcal{D}$	Sistema de ecuaciones diferenciales
$\mathcal{L}$	Función de Máxima Verosimilitud
$\mathcal{N}$	Distribución normal

2.5 Acrónimos

CIA	Criterio de Información de Akaike
CM	Cuadrados Mínimos
CMCP	Cuadrados Mínimos Cuasi Ponderado
CP	Componente Principal
CPTM	Test de las Componentes Principales para Ajustes
CO	Correntropía
Ef	Eficiencia Asintótica
ERTP	Estimador Redescendiente en Tres Partes
ESE	Error Sistemático Esporádico
ESPT	Error Sistemático que Persiste en el Tiempo
FF	Fair Function
NC	Normal Contaminada
PQ	Punto de Quiebre
RDC	Reconciliación de Datos Clásica
RDR	Reconciliación de Datos Robusta
RE	Redundancia Espacial
SEGE	Estimación Simultánea de Errores Gruesos
TG	Test Global
TIPC	Test Iterativo de las Componentes Principales

TM	Test de las Mediciones
TMIM	Test de las Mediciones Iterativo Modificado
TMP	Test de Máxima Potencia
TN	Test Nodal
TRMV	Test de Razón de Máxima Verosimilitud
UBET	Técnica de Estimación Insesgada
WE	Welsch

Capítulo 3

Reconciliación de Datos Robusta

3 Nuevas Estrategias de Reconciliación de Datos Robusta

3.1 Introducción

En este capítulo se proponen dos metodologías de Reconciliación de Datos Robusta (RDR) basadas en la función Biweight (BW), y se compara su desempeño con las estrategias que emplean los M-estimadores Welsch (WE), Cuadrados Mínimos Cuasi-Ponderados (CMCP) y Correntropía (CO). Estas tres técnicas se seleccionan con fines comparativos porque se introdujeron en la literatura de RDR durante la última década, y su desempeño ya se evaluó con respecto a sus antecesoras.

Para realizar la comparación, todos los procedimientos se sintonizan con el fin de conseguir iguales capacidades de estimación y detección de errores sistemáticos esporádicos (ESE) cuando los errores de las mediciones siguen una distribución normal estandarizada. Luego, se emplean distintos modelos de medición que simulan la presencia de ESE con el fin de evaluar el desempeño de las diferentes metodologías. Se analizan los resultados considerando métricas de desempeño relacionadas con la calidad de la estimación y las capacidades de detección e identificación de mediciones atípicas, así como también, el tiempo de cómputo.

3.2 Estimadores Robustos

Dado que la presencia de valores atípicos en las mediciones afecta las estimaciones obtenidas empleando la técnica Cuadrados Mínimos (CM), se han desarrollado estimadores robustos.

DEFINICION 3.1: Dada una distribución de probabilidad asumida o ideal de los errores de las mediciones, un estimador es robusto si resulta insensible a leves apartamientos de la distribución asumida, y es apenas menos eficiente que el estimador óptimo, obtenido empleando la función CM, cuando la suposición se satisface exactamente.

Una familia de estimadores robustos muy popular es la familia de los M-estimadores (Huber, 1964). Éstos son generalizaciones del estimador de Máxima Verosimilitud, que se introduce brevemente a continuación.

Asumamos disponer de un conjunto de N mediciones de x_i, representadas mediante la siguiente ecuación

$$y_{ip} = x_i + \varepsilon_{ip} \qquad p = 1 \ldots N \tag{3.1}$$

siendo ε_{ip} los errores de medición aleatorios. Si las observaciones son muestras de x_i independientes y obtenidas en las mismas condiciones, entonces se puede asumir que los ε_{ip} (p = 1…N) tienen la misma función de densidad de probabilidad, f_0, y son independientes. Resulta entonces, que todas las y_{ip} (p = 1…N) se distribuyen siguiendo la función de densidad de probabilidad:

$$f(y_{ip}) = f_0(y_{ip} - x_i) \tag{3.2}$$

es decir, las y_{ip} son variables aleatorias independientes e idénticamente distribuidas.

Un estimador $\hat{x}_i$ de x_i es una función de las mediciones, o sea $\hat{x}_i = \hat{x}_i(y_{i1} \ldots . y_{iN})$, y se espera obtener estimaciones tales que $\hat{x}_i \approx x_i$ con alta probabilidad. Una forma de medir esta proximidad es mediante el Error Cuadrático Medio de la estimación definido como:

$$\mathrm{ECM}(\hat{x}_i) = \mathrm{E}[(\hat{x}_i - x_i)^2] = \mathrm{Var}(\hat{x}_i) + [\mathrm{Sesgo}(\hat{x}_i, x_i)]^2 \tag{3.3}$$

donde $\mathrm{E}(\bullet)$ y $\mathrm{Var}(\bullet)$ representan los símbolos de valor esperado y varianza, respectivamente, siendo:

$$\mathrm{Var}(\hat{x}_i) = \mathrm{E}\{[(\hat{x}_i - E(\hat{x}_i)]^2\} \tag{3.4}$$

$$[\mathrm{Sesgo}(\hat{x}_i, x_i)]^2 = \mathrm{E}\{[E(\hat{x}_i) - x_i)]^2\} \tag{3.5}$$

La estimación de Máxima Verosimilitud de $\hat{x}_i$ (Canavos, 1988) se obtiene al maximizar la Función de Verosimilitud, *L*, definida como la función de densidad de probabilidad conjunta de las mediciones:

$$\mathcal{L}(y_{i1}...y_{iN}, x_i) = \prod_{p=1}^{N} f_0(y_{ip} - x_i) \tag{3.6}$$

es decir,

$$\hat{x}_i = \arg\max_{x_i} \mathcal{L}(y_{i1}...y_{iN}, x_i) \tag{3.7}$$

donde argmax significa “el valor que maximiza”.

Si f_0 se conoce exactamente, la estimación de Máxima Verosimilitud se considera óptima, ya que tiene asociada la menor varianza asintótica entre todos los estimadores no sesgados de x_i.

Si f_0 es siempre positiva, y dado que la función logaritmo es estrictamente creciente, entonces la Ec. 3.7 se puede reemplazar por:

$$\hat{x}_i = \arg\min_{x_i} \sum_{p=1}^{N} \rho(y_{ip} - x_i) \tag{3.8}$$

siendo

$$\rho = -\log f_0 \tag{3.9}$$

Si ρ es diferenciable, al derivar la Ec. 3.8 con respecto a x_i e igualar la derivada a cero resulta:

$$\sum_{p=1}^{N} \psi(y_{ip} - \hat{x}_i) = 0 \tag{3.10}$$

siendo:

$$\psi = \rho' \tag{3.11}$$

Nótese que: las raíces de la Ec. 3.10 son las estimaciones $\hat{x}_i$, y si f_0 es simétrica, entonces las funciones ρ y ψ son funciones par e impar respectivamente del ε_{ip}.

Una M-estimación de localización es una solución del Problema 3.8, que no es una estimación de Máxima Verosimilitud de ninguna distribución de probabilidad. Por lo general f_0 se conoce sólo aproximadamente, entonces los procedimientos de estimación robustos eligen una función ρ, conocida como función de pérdida o M-estimador, de manera tal que la estimación obtenida sea "casi óptima" cuando f_0 se satisface exactamente, y también cuando sólo lo hace aproximadamente (Maronna y co., 2006).

Con el fin de evaluar la calidad de las M-estimaciones, se requiere calcular sus distribuciones de probabilidad. Excepto para la media y la mediana, sólo es posible formular una aproximación de la distribución cuando la muestra tiene tamaño finito. En el Apéndice 1 se presenta una derivación heurística de esta aproximación. Si como se mencionó en capítulos anteriores, se asume que los errores aleatorios de una medición

idealmente tienen una distribución Normal con media cero y varianza σ_y^2, entonces la M-estimación de x_i, $\hat{x}_i$, se distribuye aproximadamente siguiendo una $N(x_i, v/N)$ siendo:

$$v = \frac{\mathrm{E}[\psi(\varepsilon_i)^2]}{[\mathrm{E}[\psi'(\varepsilon_i)]^2} \tag{3.12}$$

donde ψ, conocida como Función de Influencia, es la derivada del M-estimador seleccionado.

Por otra parte, se define la eficiencia asintótica de $\hat{x}_i$ como la relación

$$Ef(\hat{x}_i) = \frac{v_o}{v} \tag{3.13}$$

donde v_0 es la varianza asintótica de la estimación de Máxima Verosimilitud. La eficiencia asintótica mide cuánto se acerca la M-estimación al valor óptimo.

Una M-estimación puede considerarse como una media ponderada. En muchos casos de interés se verifica que $\psi(0) = 0$ y $\psi'(0)$ existe, de manera tal que ψ es aproximadamente lineal en el origen de coordenadas. Si se define la Función de Peso, *W*, como:

$$W(y_{ip} - \hat{x}_i) = \begin{cases} \psi(y_{ip} - \hat{x}_i)/(y_{ip} - \hat{x}_i) & \text{si } (y_{ip} - \hat{x}_i) \neq 0 \\ \psi'(0) & \text{si } (y_{ip} - \hat{x}_i) = 0 \end{cases} \tag{3.14}$$

entonces la Ec. 3.10 puede reformularse como:

$$\sum_{p=1}^{N} W(y_{ip} - \hat{x}_i)(y_{ip} - \hat{x}_i) = 0 \tag{3.15}$$

o de manera equivalente:

$$\hat{x}_i = \frac{\sum_{p=1}^{N} w_p y_{ip}}{\sum_{p=1}^{N} w_p} \tag{3.16}$$

siendo:

$$w_p = W(y_{ip} - \hat{x}_i) \tag{3.17}$$

La Ec. 3.16 expresa la M-estimación como una media ponderada de las mediciones. Dado que, en general, *W* es una función no incremental del $|(y_{ip} - \hat{x}_i)|$, las observaciones atípicas tienen asociados pesos más pequeños. Dado que los w_p dependen de $\hat{x}_i$, la Ec. 3.16 se resuelve de manera iterativa.

Los M-estimadores se clasifican en tres categorías:

- Estimadores monótonos: sus ρ son convexas y, por lo tanto no acotadas; ψ es una función creciente de a. Un ejemplo de M-estimadores monótonos es la Función de Huber (HU):

$$\rho_{HU} = \begin{cases} a^2 & |a| \le c_{HU} \\ 2c_{HU}|a| - {c_{HU}}^2 & |a| > c_{HU} \end{cases} \tag{3.18}$$

$$\psi_{HU} = \begin{cases} a & |a| \le c_{HU} \\ \text{sgn}(a)c_{HU} & |a| > c_{HU} \end{cases} \tag{3.19}$$

$$\text{sgn}(a) = \begin{cases} -1 & a < 0 \\ 0 & a = 0 \\ 1 & a > 0 \end{cases} \tag{3.20}$$

$$W_{HU} = \begin{cases} 1 & |a| \le c_{HU} \\ \dfrac{\text{sgn}(a)c_{HU}}{a} & |a| > c_{HU} \end{cases} \tag{3.21}$$

donde c_{HU} es una constante que regula la *Ef* del estimador (Huber, 1964). Otros ejemplos de este tipo de M-estimadores son: la Fair Function (Rey, 1983), CMCP (Zhang y co., 2010) y Logística (Özyurt y Pike, 2004).

- Estimadores redescendentes con ρ no acotada: sus ψ tiende a cero en el infinito. Un ejemplo de este tipo de M-estimadores es la Función de Cauchy (Özyurt y Pike, 2004):

$$\rho_C = \frac{{c_C}^2}{2}\log\left[1+\left(\frac{a}{c_C}\right)^2\right] \tag{3.22}$$

$$\psi_C = \frac{a}{1+\frac{a^2}{c_C}} \tag{3.23}$$

$$W_C = \frac{1}{1+\frac{a^2}{c_C}} \tag{3.24}$$

donde c_C es un parámetro que fija la *Ef* del estimador. A este grupo pertenecen también las funciones de Lorentz (Huber, 1981), WE (Rey, 1983), y CO (Chen y co., 2013).

- Estimadores redescendentes con ρ acotada: las ρ se definen a tramos y las ψ son iguales a cero a partir de un dado valor de *a*. Un ejemplo de este tipo de M-estimadores es la BW (Rey y co., 1983)

$$\rho_{BW} = \begin{cases} 1-\left[1-(a/c_{BW})^2\right]^3 & if \quad |a| \le c_{BW} \\ 1 & if \quad |a| > c_{BW} \end{cases} \tag{3.25}$$

$$\psi_{BW} = \begin{cases} a\left[1-\left(\frac{a}{c_{BW}}\right)^2\right]^2 & if \quad |a| \le c_{BW} \\ 0 & if \quad |a| > c_{BW} \end{cases} \tag{3.26}$$

$$W_{BW} = \begin{cases} \left[1-\left(\frac{a}{c_{BW}}\right)^2\right]^2 & if \quad |a| \le c_{BW} \\ 0 & if \quad |a| > c_{BW} \end{cases} \tag{3.27}$$

donde c_{BW} se ajusta para conseguir un valor de *Ef* dado. La función de Hampel (HA) es otro M-estimador perteneciente a esta clase (Arora y Biegler, 2001)

La comparación de estas tres clases de estimadores indica que:

- Los M-estimadores monótonos son sensibles a valores atípicos grandes, porque los pesos correspondientes a éstos son mayores que los proporcionados por los M-estimadores redescendentes. Por tal motivo, también pueden tener baja eficiencia para distribuciones de errores de colas pesadas. Además, cabe notar que la solución de la Ec. 3.8 tiene un solo mínimo local, por lo tanto, el valor empleado para comenzar el proceso de iteración influye en el número de iteraciones pero no en el valor de la solución final.
- Los M-estimadores redescendentes otorgan bajos pesos a los valores atípicos, y por lo tanto son más robustos que los monótonos. También son más eficientes para distribuciones de errores de colas pesadas, pero el problema de optimización 3.8 puede tener varios mínimos locales, por lo que se requiere un punto inicial adecuado para asegurar la obtención de una buena solución. En especial, los M-estimadores redescendentes con ρ acotada rechazan completamente los valores atípicos grandes, y con una adecuada elección de sus parámetros pueden alcanzar una alta eficiencia tanto

para la distribución de error Normal como para las distribuciones de colas pesadas (Maronna y co., 2006).

Este análisis muestra que: los M-estimadores redescendentes con ρ acotada resultan atractivos debido a su alta robustez a valores atípicos de gran magnitud, mientras que los monótonos alcanzan el óptimo local independientemente del punto inicial.

3.3 Formulación del Problema de Reconciliación de Datos Robusta

Teniendo en cuenta los conceptos previamente introducidos, se definirá como una M-estimación de localización del estado del proceso en el tiempo *j* tal que satisface su modelo operativo a la solución del siguiente problema de optimización:

$$\begin{aligned} [\hat{\mathbf{x}}_j^R, \hat{\mathbf{u}}_j^R] = \underset{x_j, u_j}{Min} \quad & \sum_{p=j-N+1}^{j} \sum_{i=1}^{I} \rho(a_{ip}) \\ st. \\ & \mathbf{f}(\mathbf{x},\mathbf{u})=\mathbf{0} \\ & \mathbf{h}(\mathbf{x},\mathbf{u}) \le \mathbf{0} \\ & \mathbf{x}^{\mathbf{L}} \le \mathbf{x} \le \mathbf{x}^{\mathbf{U}} \\ & \mathbf{u}^{\mathbf{L}} \le \mathbf{u} \le \mathbf{u}^{\mathbf{U}} \end{aligned} , \tag{3.28}$$

donde

$$a_{ip} = \frac{y_{ip} - x_{ij}}{\sigma_{y,i}} , \tag{3.29}$$

representa el ajuste estandarizado de la *i*-ésima variable medida. El estado del proceso en el intervalo de muestreo *j* queda definido por las estimaciones de los vectores **x**, de dimensión *I*, y **u**, de dimensión *U*, correspondientes a las variables medidas y no medidas respectivamente, que satisfacen el modelo del proceso en dicho tiempo, $[\hat{\mathbf{x}}_j^R, \hat{\mathbf{u}}_j^R]$. Éste se representa mediante sistemas de ecuaciones algebraicas de igualdad **f**, de desigualdad

h y comprende límites para las variables. Si se denota como $\mathbf{y}_p = (y_{1p}, \ldots, y_{Ip})$ al vector de observaciones en el intervalo de muestreo *p*, y el proceso se observa para un horizonte de datos de tamaño *N*, entonces, la M-estimación de localización en el intervalo de muestreo *j* se basa en la información contenida en los vectores de observación $\mathbf{y}_{j-N+1}, \ldots, \mathbf{y}_j$. Es decir, la estimación se beneficia empleando tanto la redundancia temporal de cada variable medida como la redundancia espacial provista por las ecuaciones que las relacionan.

3.4 Estrategias de Reconciliación Robusta de Datos Propuestas en la Última Década

Durante la última década se utilizó la función WE para resolver problemas de RDR y se formularon las funciones de pérdida CMCP y CO con igual propósito. Las estrategias basadas en dichas funciones se emplean con fines de comparación en esta tesis y se las describe brevemente a continuación.

3.4.1 M-estimador de Welsch

El M-estimador WE, introducido por Dennis y Welsch (1976), es un estimador redescendente con ρ no acotada, que presenta redescendencia suave. Sus ρ, ψ y W se presentan a continuación:

$$\rho_{WE} = c_{WE}^2 \left\{ 1 - \exp\left[-\left(\frac{a}{c_{WE}} \right)^2 \right] \right\} \tag{3.30}$$

$$\psi_{WE} = a\, e^{-\left(\frac{a}{c_{WE}} \right)^2} \tag{3.31}$$

$$W_{WE} = e^{-\left(\frac{a}{c_{WE}}\right)^2}$$

(3.32)

siendo c_{WE} el parámetro del M-estimador. Su ψ se aproxima asintóticamente a cero para valores grandes de *a*.

Esta función sólo se utilizó para reconciliar las mediciones contenidas en ventanas móviles de datos correspondientes a procesos dinámicos (Martinez Prata y co., 2008; 2010). El parámetro se ajustó para satisfacer una dada *Ef*. Para inicializar el problema de estimación, se usaron los valores de las mediciones como valores iniciales de las variables independientes.

3.4.2 M-estimator Cuadrados Mínimos Cuasi Ponderado

Las ρ, ψ y *W* del M-estimador CMCP, propuesto por Zhang y co. (2010), se definen como sigue:

$$\rho_{CMCP} = \frac{a^2}{1 + c_{CMCP}\,|a|} \tag{3.33}$$

$$\psi_{CMCP} = \begin{cases} \dfrac{4a - c_{CMCP}a^2}{\left(2 + c_{CMCP}a\right)^2} & a < 0 \\ \dfrac{4a + c_{CMCP}a^2}{\left(2 + c_{CMCP}a\right)^2} & a \geq 0 \end{cases} \tag{3.34}$$

$$W_{CMCP} = \begin{cases} \dfrac{4 - c_{CMCP}a}{\left(2 + c_{CMCP}a\right)^2} & a < 0 \\ \dfrac{4 + c_{CMCP}a}{\left(2 + c_{CMCP}a\right)^2} & a \geq 0 \end{cases} \tag{3.35}$$

La adición del término $c_{CMCP}|a|$ al denominador de la función CM reduce el efecto de las mediciones atípicas con valores grandes, siendo c_{CMCP} un parámetro. Esta función es un estimador monótono y su $\psi \to 1/c_{CMQP}$ cuando $a \to \infty$.

El estimador CMCP se utilizó para resolver problemas de RDR en procesos que operan en estado estacionario y cuya operación se representa empleando sistemas lineales de ecuaciones. El parámetro se ajustó minimizando el Criterio de Información de Akaike. No se proporcionó ninguna discusión sobre la selección del punto inicial del problema de estimación ni de sus requerimientos de tiempo de cómputo.

3.4.3 M-estimador Correntropía

Las ρ, ψ y W del M-estimador CO, introducido por Chen y co. (2013), se presentan a continuación:

$$\rho_{CO} = \frac{1}{c_{CO}\sqrt{2\pi}} \exp\left[-\left(\frac{a^2}{2a_{CO}{}^2}\right)\right] \tag{3.36}$$

$$\psi_{CO} = \frac{a\, e^{-\left(\frac{a^2}{2c_{co}^2}\right)}}{\sqrt{2\pi}c_{co}^3} \tag{3.37}$$

$$W_{CO} = \frac{e^{-\left(\frac{a^2}{2c_{co}^2}\right)}}{\sqrt{2\pi}c_{co}^3} \tag{3.38}$$

La función del Kernel gaussiano depende de su ancho de banda c_{CO}. La ρ del M-estimador CO tiende rápidamente a cero para $|a| > c_{CO}$.

Se resolvieron problemas de RDR para procesos que operan en estado estacionario y cuya operación se representa mediante conjuntos de ecuaciones algebraicas lineales y no lineales. El problema de estimación se inicializó con la solución obtenida aplicando la función CM como estimador. Se propuso formular primero el problema de optimización sin restricciones (Romagnoli y Sánchez, 2000) y resolverlo utilizando un procedimiento iterativo.

3.5 Nuevas estrategias de Reconciliación Robusta de Datos

A continuación se presentan nuevas metodologías de RDR desarrolladas bajo la premisa de combinar las ventajas de los M-estimadores monótomos y redescendentes resaltadas en la Sección 3.2. Las nuevas estrategias se denominan Método Simple (MSi) y Método Sofisticado (MSo).

El estimador monótono seleccionado es el de HU, que ha sido el más usado dentro de este tipo de M-estimadores, aunque también pueden emplearse otros de la misma clase. En relación con los estimadores redescendentes, las funciones BW y HA, comprendidas en el grupo de M-estimadores redescendentes con ρ acotada, resultan atractivas para la RDR debido a que permiten eliminar completamente el efecto de los errores sistemáticos de gran magnitud. La comparación de los tiempos de cómputo necesarios para que estas funciones alcancen el óptimo dio resultados favorables para la BW (Sánchez y Maronna, 2009), por lo cual se selecciona este M-estimador para el desarrollo de metodologías robustas (Llanos y co., 2015).

3.5.1 Método Simple

El MSi comprende los dos pasos siguientes:

Paso 1: Cálculo de una M-estimación de localización de la *i*-ésima variable medida (*i* = 1: *I*) en el tiempo *j*, $\tilde{y}_{ij}$, empleando las mediciones obtenidas en un horizonte de tiempo de tamaño *N*. Se emplea como M-estimador la función BW.

$$\tilde{y}_{ij} = Min \quad \sum_{p=j-N+1}^{j} \rho_{BW}\left(\frac{y_{ip} - \tilde{y}_i}{\sigma_i}\right) \tag{3.39}$$

Se obtiene así una estimación inicial de la *i*-ésima variable, $\tilde{y}_{ij}$, que es la mediana robusta de $\{y_{i,j-N+1},....,y_{i,j}\}$. El aprovechamiento de la redundancia temporal provista por las observaciones repetidas es de gran utilidad para compensar la escasez de redundancia espacial, que comúnmente se observa en los procesos reales (Maronna y Arcas, 2009).

Paso 2: Cálculo de una M-estimación del estado del proceso en el tiempo *j*, que satisfaga el modelo que lo representa, empleando el M-estimador de Huber. Dicha estimación es la solución del siguiente problema de optimización:

$$[\hat{\mathbf{x}}_j^R, \hat{\mathbf{u}}_j^R] = \underset{x_j, u_j}{Min} \quad \sum_{i=1}^{I} \rho_{HU}\left(\frac{\tilde{y}_i - x_i}{\sigma_i}\right)$$

$$st.$$

$$\begin{aligned} &\mathbf{f(x,u)=0} \\ &\mathbf{h(x,u)} \le \mathbf{0} \\ &\mathbf{x^L} \le \mathbf{x} \le \mathbf{x^U} \\ &\mathbf{u^L} \le \mathbf{u} \le \mathbf{u^U} \end{aligned} \tag{3.40}$$

Estos dos pasos trabajan conjuntamente. El primero ayuda a reducir el efecto de los valores atípicos que pueden sesgar la estimación cuando se utiliza un M-estimador

monótono para resolver el problema 3.28. Por otra parte, la resolución del problema 3.40 es más sencilla, en comparación con el procedimiento que emplea un M-estimador redescendente, porque su solución es única. Así, los valores utilizados para iniciar el proceso iterativo pueden influir en el número de iteraciones pero no en el resultado final. La solución secuencial de los Problemas 3.39 y 3.40 no coincide exactamente con la del Problema 3.28, pero a los fines prácticos la diferencia es poco significativa.

3.5.2 Método Sofisticado

El MSo incorpora el siguiente paso al MSi:

Paso 3: Se resuelve el Problema 3.28 empleando la función BW y la solución del Problema 3.40 como M-estimador y punto inicial, respectivamente, es decir, la M-estimación del estado del sistema es la solución de:

$$[\hat{\mathbf{x}}_j^R,\ \hat{\mathbf{u}}_j^R] = \underset{x_j,u_j}{Min} \sum_{p=j-N+1}^{j} \sum_{i=1}^{I} \rho_{BW}(a_{ip})$$

$$st.$$

$$\begin{aligned} \mathbf{f(x,u)} &= \mathbf{0} \\ \mathbf{h(x,u)} &\le \mathbf{0} \\ \mathbf{x^L} \le \mathbf{x} &\le \mathbf{x^U} \\ \mathbf{u^L} \le \mathbf{u} &\le \mathbf{u^U} \end{aligned} \qquad (3.41)$$

y se la considera como la estimación final.

El Paso 3 es una etapa de refinamiento de la solución obtenida en el Paso 2. Dado que la función BW es un estimador redescendente con ρ acotada; el punto inicial del problema de optimización debe ser robusto para asegurar la convergencia a una buena solución. En 3.41 se utilizan todas las mediciones del horizonte de tiempo para encontrar la solución que satisface el modelo del proceso.

3.6 Análisis de Desempeño

En esta sección se introduce la metodología empleada para comparar el desempeño de las estrategias de RDR propuestas en relación con el obtenido al aplicar las metodologías que fueron presentadas en la última década.

En esta tesis se analizan las capacidades de las estrategias seleccionadas para estimar variables y detectar e identificar ESE en dos casos de estudio. Los resultados de los procedimientos se examinan para tres modelos de medición diferentes:

Modelo 1: mediciones sin valores atípicos. Se considera que los errores aleatorios estandarizados se distribuyen siguiendo la misma f_0, y se asume que ésta es la N (0,1);

Modelo 2: mediciones con errores sistemáticos esporádicos (ESE). El error de la observación se representa empleando una f_0 simétrica de colas pesadas. Se selecciona una distribución normal contaminada $f_0 \sim (1-\xi_c)N(0,1)+\xi_c N(0,R^2)$ donde ξ_c denota la tasa de contaminación. Es decir, con probabilidad ξ_c un error aleatorio estandarizado normal se multiplica por una constante R.

Modelo 3: mediciones con fallas aleatorias. La mayoría de las observaciones siguen el primer modelo, pero aleatoriamente una proporción ξ_c no lo obedece. Entre los posibles escenarios de falla, el error de la medición se representa como un valor fijo $K\sigma_i$.

Las estimaciones no tienen sesgo cuando las mediciones se ajustan a los dos primeros modelos. Por lo tanto, el ECM refleja sólo la varianza de las estimaciones; es deseable que éstas tengan una alta eficiencia para ambos modelos. Por el contrario, el ECM refleja tanto la varianza como el sesgo de la estimación para el tercer modelo, y ambos deben controlarse.

Para comparar las capacidades de estimación de las diferentes técnicas, la *Ef* de los diferentes M-estimaciones se fija en 0,95 mediante el ajuste adecuado de sus parámetros. Özyurt y Pike (2004) trataron esta cuestión de la misma manera. Para los estimadores CMCP y CO la sintonización se realiza utilizando el procedimiento Jackknife (Rey, 1983). La Tabla 3.1 presenta los valores de los parámetros para cada ρ.

Tabla 3.1 Parámetros de ajuste para *Ef*= 0,95

c_{CO}	c_{CMCP}	c_{WE}	c_{BW}	c_{HU}
2,05	0,89	2,98	4,68	1,37

Las estrategias de tipo adaptativo utilizan información de cada muestra del proceso con el fin de optimizar los parámetros del estimador. Existen diferentes maneras de abordar este tema, por ejemplo: minimizar una estimación de la varianza, maximizar la Función de Verosimilitud T Generalizada evaluada en las estimaciones iniciales de los estados del proceso, minimizar el criterio de información de Akaike, etc. Sin embargo, un M-estimador adaptativo no es necesariamente mejor que un M-estimador debidamente sintonizado. Las siguientes evidencias apoyan esta afirmación (Sánchez y Maronna, 2009):

- Numerosas simulaciones de la Estadística Robusta (Maronna y co., 2006) han demostrado que los M-estimadores adaptativos, a pesar de su mayor complejidad computacional, no mejoran el correcto desempeño de los M-estimadores con parámetros fijos elegidos idóneamente.
- Debe recordarse que el valor verdadero del parámetro óptimo es desconocido para los estimadores adaptativos. Sólo está disponible una estimación del mismo que tiene un cierto sesgo y varianza, los que a su vez se propagan a las estimaciones obtenidas

mediante la RDR. Por lo tanto, este enfoque es confiable sólo para tamaños de muestra muy grandes. Recordemos que el ECM de un estimador puede descomponerse como indica la Ec. 3.3. Típicamente, la varianza disminuye con 1 / *N*, pero esto no sucede para el sesgo ocasionado por la contaminación del tercer modelo. Por lo tanto, para tamaños de muestras grandes, donde los estimadores adaptativos tienen sentido, la eficiencia no es tan importante como el sesgo.

En la Fig. 3.1 se representan las funciones HU, BW, WE, CMCP y CO luego de ajustar sus parámetros para que todos los M-estimadores produzcan estimaciones con una *Ef* = 095. En las Fig. 3. 2 y Fig. 3.3, se grafican las funciones ψ y *W* de dichos estimadores, respectivamente.

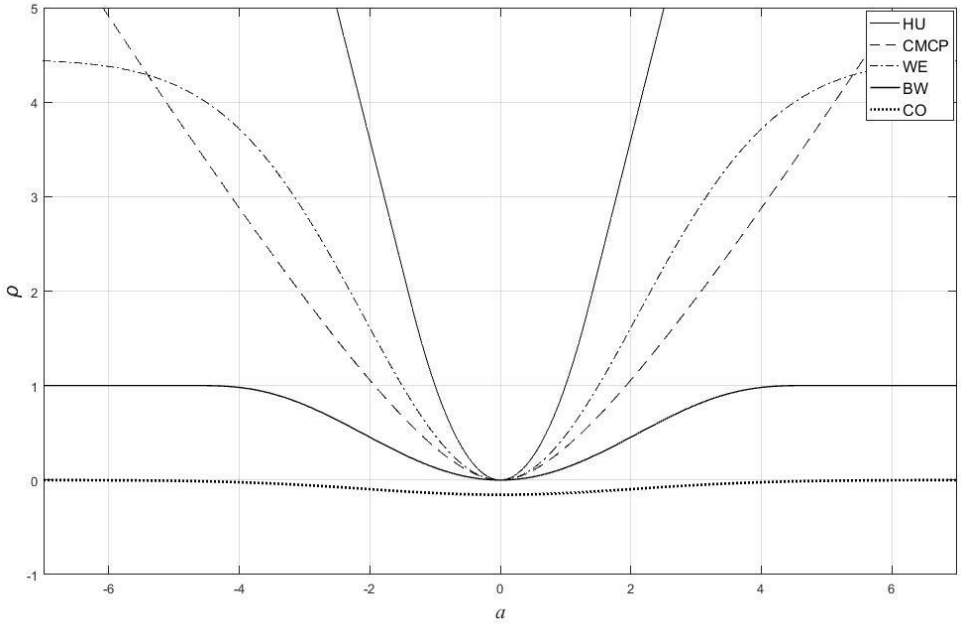


Figura 3.1 Funciones de pérdida de los M-estimadores

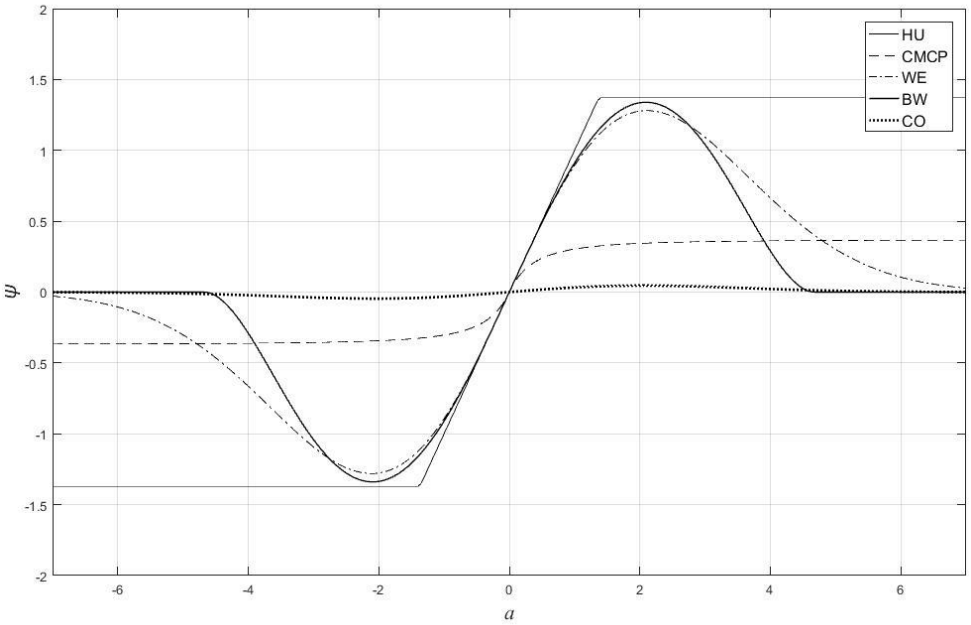


Figura 3.2 Funciones de Influencia de los M-estimadores

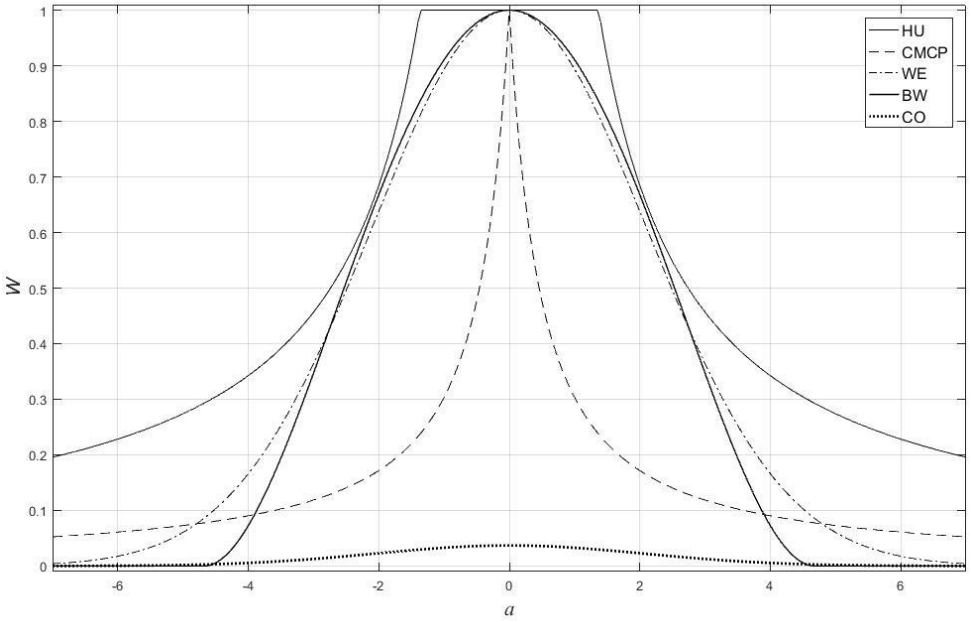


Figura 3.3 Funciones de Peso de los M-estimadores

Una vez seleccionado un cierto modelo para las mediciones, se realizan 10000 simulaciones del procedimiento de RDR, para un horizonte de tiempo N=10, con el fin de calcular las medidas de desempeño de las cinco metodologías analizadas. Las métricas

utilizadas para tal fin son: ECM, AVTI (Número Promedio de Errores Tipo I) y OP (Desempeño Global), las cuales fueron propuestas por Narasimhan y Mah (1987) y se definen a continuación:

$$\text{ECM}=\frac{1}{I\ Ns}\sum_{k=1}^{Ns}\sum_{i=1}^{I}\left(\frac{\hat{x}_i - x_i}{\sigma_{y,i}}\right)^2 \tag{3.42}$$

$$\text{AVTI} = \frac{\#(\text{ESE incorrectamente identificados})}{Ns} \tag{3.43}$$

$$\text{OP} = \frac{\#(\text{ESE correctamente identificados})}{\#(\text{ESE simulados})} \tag{3.44}$$

siendo *Ns* el número de simulaciones.

El punto de corte de cada técnica, es decir el valor más allá del cual las mediciones se consideran valores atípicos, se ajusta por prueba y error de tal manera que el AVTI sea aproximadamente 0,05 cuando no hay valores atípicos y las mediciones se generen usando una distribución Normal (Modelo 1). Esta práctica proviene de los primeros trabajos de Reconciliación de Datos, RD, (Iordache y co., 1985), y garantiza que todos los procedimientos tienen el mismo comportamiento cuando no hay valores atípicos.

En cuanto al punto inicial del problema de optimización, Chen y co. (2013) informaron que inicializaron el M-estimador CO utilizando la solución del procedimiento que emplea la función CM. Por el contrario, Zhang y co. (2010) no hicieron referencia a esta cuestión cuando aplicaron la función CMCP. Dado que Chen y co. (2013) compararon las estimaciones obtenidas usando las funciones CO y CMCP para el proceso conocido como Red de Ingreso de Vapor (*Steam Metering Network*, SMN, Serth y Heenan, 1986), en este trabajo de tesis se asume la inicialización utilizada por Chen y co. (2013) para ambos procedimientos. Además, se utiliza el mismo punto de partida para

resolver el problema de RDR empleando la función WE, porque tampoco aparece ninguna mención en la literatura sobre este aspecto.

Los procedimientos se ejecutaron utilizando un Procesador Intel ® Core (TM) i7 CPU 930 @ 2,80 GHz, 8 GB de RAM, utilizando el código de programación cuadrática sucesiva de MatLab Release 7.12 (R2011a).

3.7 Resultados

A continuación se presentan y analizan en detalle los resultados obtenidos para dos casos de estudio.

3.7.1 Red de Ingreso de Vapor (SMN)

La SMN involucra 28 corrientes que interconectan 11 unidades (Fig. 3.4). Se miden los caudales de todas las corrientes. Inicialmente se considera el Modelo 1 de las mediciones, es decir, éstas no presentan valores atípicos. Se generan errores aleatorios considerando que las desviaciones estándar de las observaciones son 2,5% de sus valores reales. Las Tabla 3.2 y la Tabla 3.3 presentan los puntos de corte de las metodologías y las medidas de desempeño, respectivamente.

Tabla 3.2 Puntos de Corte - SMN

MSi	MSo	CO	CMCP	WE
3.84	3.8416	3.8165	3.753	3.811

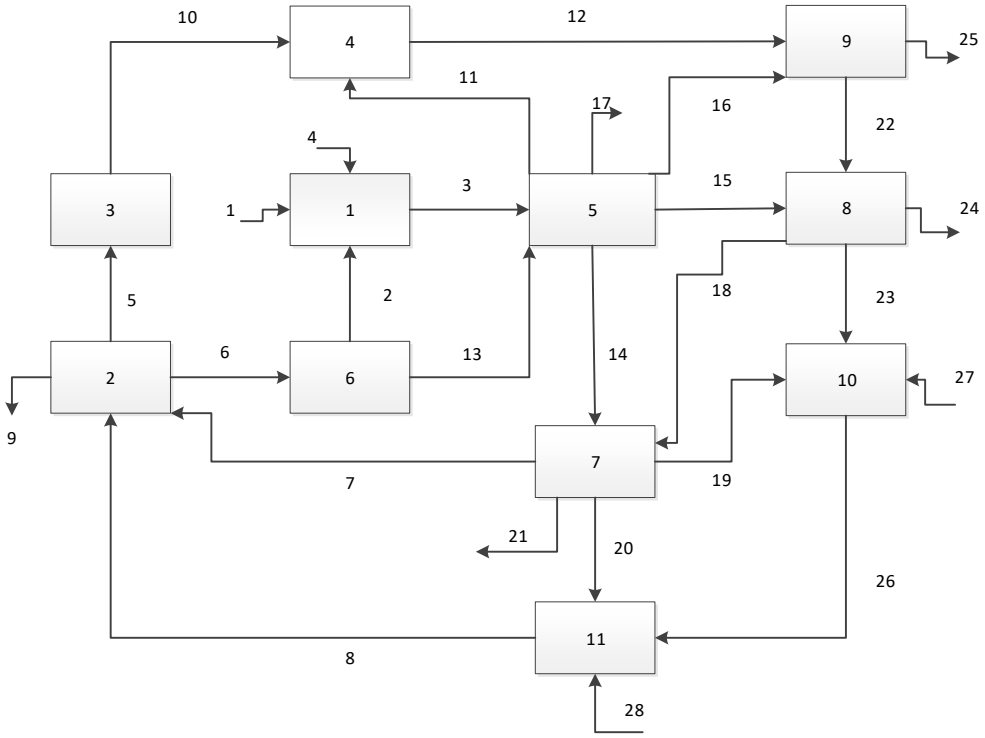


Figura 3.4 Red de Ingreso de Vapor

Tabla 3.3 Resultados para el Modelo 1 - SMN

AVTI					ECM * 10^2				
MSi	MSo	CO	CMCP	WE	MSi	MSo	CO	CMCP	WE
0.0499	0.0499	0.05	0.0499	0.0499	6.384	6.387	6.4129	6.4109	6.3796

La Tabla 3.4 muestra el desempeño de las metodologías, para diferentes valores de *R*, cuando se aplica el Modelo 2. Sólo los resultados del AVTI y ECM se grafican en la Fig. 3.5 porque los OP son similares para todas las estrategias. En la Tabla 3.5, se presenta el promedio de los tiempos de ejecución para las 10000 simulaciones.

En la Fig. 3.5 se puede observar que:

- Los resultados de las estrategias MSi, MSo, WE y CO son similares para $R \in [2, 10]$, pero MSi y MSo superan a CO y WE para $R > 10$;

- Los comportamientos de CO y WE son comparables para $R > 10$, aunque hay evidencia de una ligera superioridad de WE con respecto a CO;
- Los valores de AVTI y ECM obtenidos para la CMCP aumentan con R y son más pobres que los proporcionados por MSi y MSo, excepto para $R = 2$. En contraste, CMCP se comporta mejor que WE y CO para grandes contaminaciones;
- Los valores de desempeño de MSi y MSo son afectados sólo ligeramente por el valor de R.

Tabla 3.4 Resultados para el Modelo 2- SMN

	R	2	5	10	14	15	18	20
	MSi	0.052	0.053	0.049	0.051	0.048	0.049	0.046
	MSo	0.053	0.052	0.049	0.050	0.048	0.048	0.046
AVTI	CO	0.050	0.051	0.050	0.063	0.068	0.102	0.126
	CMCP	0.048	0.054	0.055	0.058	0.060	0.056	0.061
	WE	0.050	0.051	0.050	0.060	0.065	0.097	0.117
	MSi	0.055	0.438	0.698	0.779	0.791	0.817	0.825
	MSo	0.055	0.438	0.698	0.779	0.791	0.817	0.825
OP	CO	0.055	0.438	0.698	0.778	0.791	0.816	0.824
	CMCP	0.055	0.437	0.697	0.778	0.791	0.816	0.824
	WE	0.055	0.438	0.698	0.778	0.791	0.816	0.824
	MSi	7.531	7.921	7.643	7.518	7.473	7.445	7.411
	MSo	7.509	7.871	7.606	7.485	7.446	7.416	7.393
ECM*10^2	CO	7.489	7.922	7.699	8.197	8.555	12.547	20.678
	CMCP	7.510	9.223	10.322	10.697	10.822	10.925	10.997
	WE	7.458	7.917	7.684	8.097	8.426	12.425	17.808

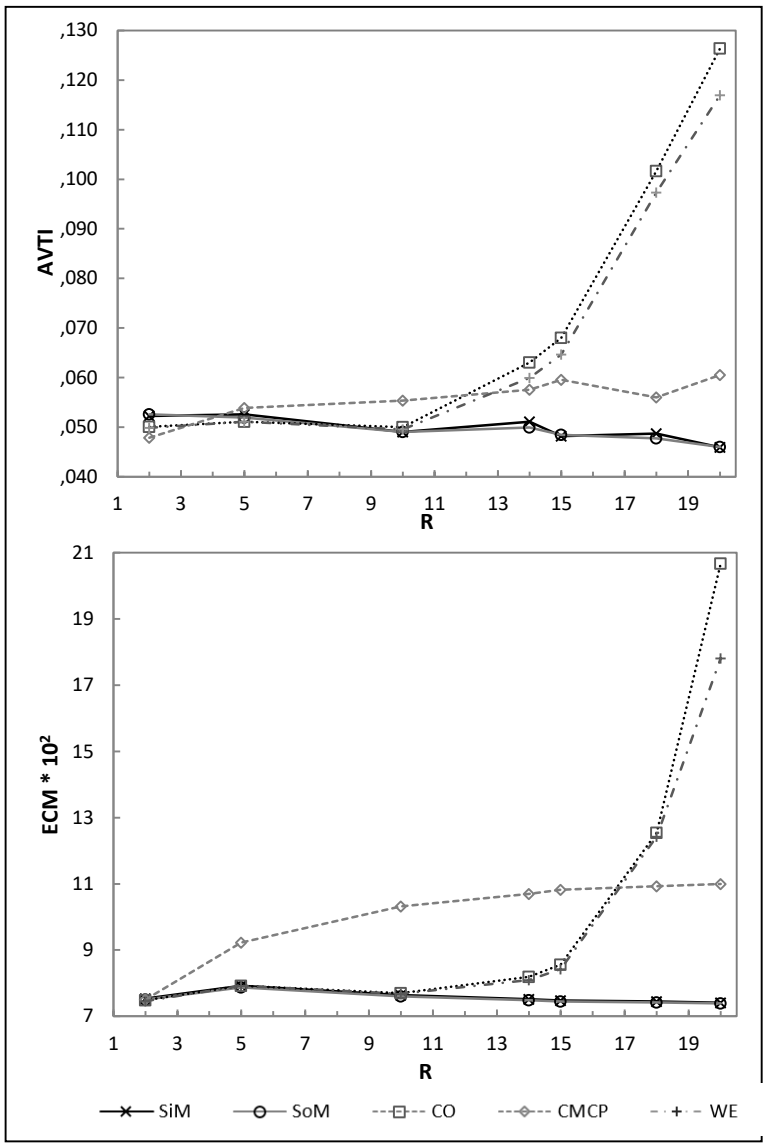


Figura 3.5 AVTI y ECM para el Modelo 2 - SMN

El análisis de las Tablas 3.4 y 3.5 indica que MSo proporciona valores de OP y ECM ligeramente mejores que los obtenidos con MSi, a expensas de un incremento en el tiempo de cómputo ocasionado por la ejecución del Paso3.

Tabla 3.5 Promedio de Tiempos de Ejecución (s) para el Modelo 2 - SMN

MSi	MSo	CO	CMCP	WE
147.8	374.0	302.2	274.0	227.8

Los resultados anteriores señalan que MSi funciona bien para la estimación de variables y la detección de valores atípicos para todo el rango de valores de *R*, y sus requerimientos de tiempo de cómputo son los más bajos.

A continuación, se presentan los resultados alcanzados cuando se aplica el Modelo 3 para las mediciones. La Tabla 3.6 contiene los valores de las medidas de desempeño para diferentes *K*. Además, se grafican el AVTI y ECM en la Fig. 3.6, y los tiempos de ejecución promedio se muestran en la Tabla 3.7.

Si las mediciones no obedecen la distribución normal contaminada, puede observarse en la Fig. 3.6 que:

- Los valores de AVTI y ECM para el CMCP aumentan con *K*;
- Los índices de desempeño para CO y WE son comparables para todos los valores de *K*;
- Los valores de AVTI de MSi y MSo son mayores que los correspondientes a WE y CO para *K* [1,..., 4]; MSi y MSo tienden al comportamiento de WE y CO para $K \geq 5$;
- El ECM obtenido usando MSo es menor que el alcanzado por WE y CO para $K = 1$, 4-8, y ligeramente mejor que el obtenido usando MSi.

A partir del análisis de los valores de OP presentados en la Tabla 3.6, se puede concluir que el comportamiento de todas las técnicas es similar, excepto para $K = 4$. En este caso, las metodologías basadas en la función BW presentan valores de OP más altos. Respecto a los tiempos de cómputo promedio, la Tabla 3.7 muestra que los requerimientos computacionales de MSo son los más grandes y lo contrario sucede con los de MSi.

Tabla 3.6 Resultados para el Modelo 3 - SMN

	K	1	2	3	4	5	6	7	8	9	10
AVTI	MSi	0.049	0.069	0.091	0.092	0.057	0.048	0.048	0.048	0.048	0.048
	MSo	0.049	0.069	0.088	0.083	0.056	0.048	0.048	0.048	0.048	0.048
	CO	0.049	0.067	0.080	0.073	0.056	0.048	0.047	0.047	0.047	0.047
	CMCP	0.047	0.061	0.072	0.082	0.090	0.097	0.101	0.106	0.109	0.111
	WE	0.049	0.067	0.080	0.074	0.057	0.048	0.047	0.047	0.047	0.047
OP	MSi	0.000	0.000	0.002	0.584	0.998	1.000	1.000	1.000	1.000	1.000
	MSo	0.000	0.000	0.002	0.607	0.998	1.000	1.000	1.000	1.000	1.000
	CO	0.000	0.000	0.001	0.541	0.996	1.000	1.000	1.000	1.000	1.000
	CMCP	0.000	0.000	0.000	0.384	0.971	0.999	1.000	1.000	1.000	1.000
	WE	0.000	0.000	0.001	0.528	0.995	1.000	1.000	1.000	1.000	1.000
$ECM*10^2$	MSi	7.681	12.539	15.092	11.601	7.545	7.167	7.166	7.166	7.166	7.166
	MSo	7.738	12.522	14.491	10.863	7.437	7.153	7.153	7.153	7.153	7.153
	CO	7.850	12.478	13.969	11.176	8.495	7.527	7.273	7.201	7.184	7.181
	CMCP	8.150	11.784	14.363	16.081	17.234	18.028	18.590	19.000	19.305	19.539
	WE	7.782	12.406	14.160	11.545	8.677	7.565	7.261	7.172	7.149	7.144

Tabla 3.7 Promedio de Tiempos de Ejecución (s) para el Modelo 3 - SMN

MSi	MSo	CO	CMCP	WE
132.3	356.8	314.2	274.3	231.0

La Figura 3.6 muestra que el AVTI y ECM de todos los M-estimadores redescendentes cambian con el incremento de *K* de manera similar para el Modelo 3. Si se consideran las medidas de desempeño, no hay evidencia de una clara superioridad entre las metodologías analizadas. Se puede observar que la técnica MSi proporciona un buen equilibrio entre las capacidades de detección e identificación de valores atípicos y la carga computacional del procedimiento.

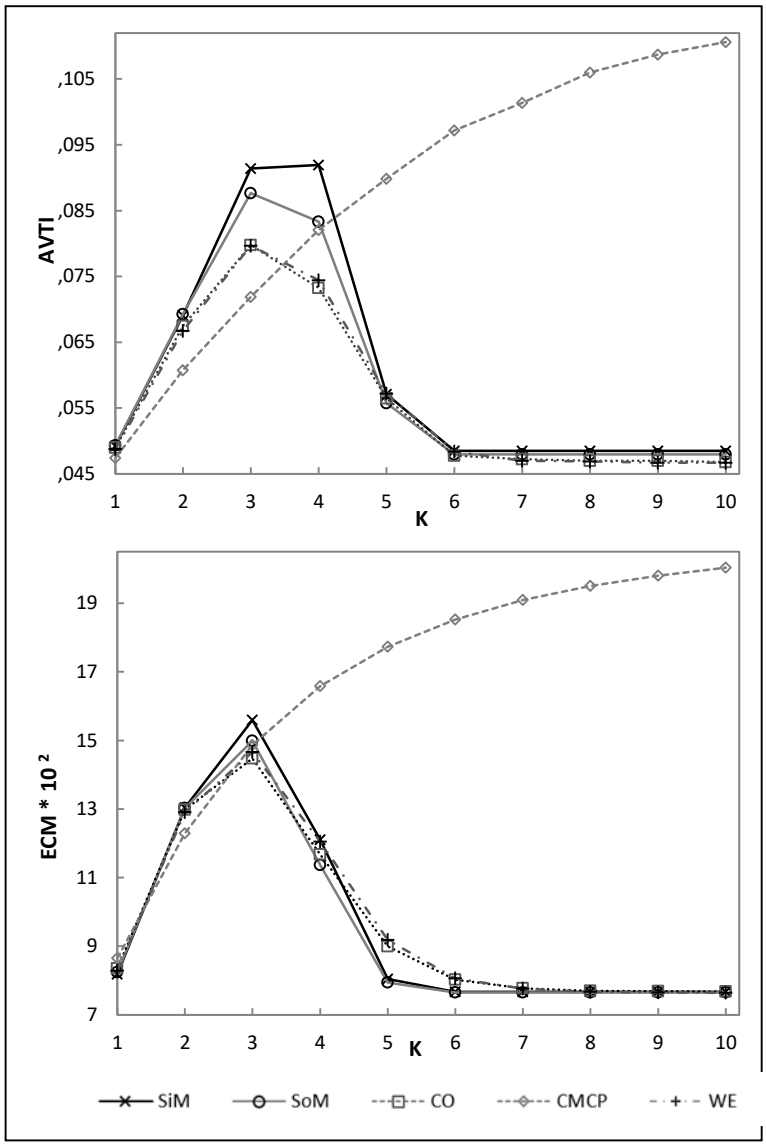


Figura 3.6 AVTI y ECM para el Modelo 3 - SMN

Los siguientes comentarios surgen del análisis de los resultados anteriores:

- Aunque las estrategias basadas en las funciones CMCP, CO y WE tienen la misma inicialización, el comportamiento de la técnica CMCP es diferente de los otros dos métodos debido a que la función CMCP es un M-estimador monótono;
- Tanto las funciones WE como CO son M-estimadores redescendentes que comprenden funciones exponenciales. Por lo tanto, se comportan de manera similar;

- La función BW rechaza los valores atípicos si el ajuste estandarizado de la observación es mayor que 4,68, por lo que las medidas de desempeño de MSi y MSo no cambian significativamente para K> 5.

3.7.2 Ejemplo No lineal

El segundo caso de estudio se extrae del artículo de Pai y Fisher (1988), y se simboliza como P & F. Comprende seis restricciones de igualdad no lineales, que se definen en términos de cinco variables redundantes medidas y tres variables no medidas observables, las cuales se presentan a continuación:

$$Eq.1 = 0.5x_1^2 - 0.7x_2 + x_3u_1 + x^2u_1u_2 + 2x_3u_3^2 - 255.8 = 0$$
$$Eq.2 = x_1 - 2x_2 + 3x_1x_3 - 2x_2u_1 - x_2u_2u_3 + 111.2 = 0$$
$$Eq.3 = x_3u_1 - x_1 + 3x_2 + x_1u_2 - x_3u_3^{1/2} - 33.57 = 0$$
$$Eq.4 = x_4 - x_1 - x_3^2 + u_2 + 3u_3 = 0$$
$$Eq.5 = x_5 - 2x_3u_2u_3 = 0$$
$$Eq.6 = 2x_1 + x_2x_3u_1 + u_2 - u_3 - 126.6 = 0$$

Los errores aleatorios se generan considerando las desviaciones estándar sugeridas por dichos autores. El mismo tipo de análisis realizado para el ejemplo lineal se presenta para el no lineal.

La Tabla 3.8 y la Tabla 3.9 muestran los puntos de corte de las metodologías y las medidas de desempeño, respectivamente, para el Modelo 1.

Con respecto al Modelo 2, la Tabla 3.10 contiene los índices de desempeño para diferentes valores de R, mientras que la Fig. 3.7 sólo muestra los registros del AVTI y ECM porque los OP obtenidos son similares para todas las estrategias. En la Tabla 3.11, se informa el promedio de los tiempos de ejecución para las 10000 simulaciones.

Tabla 3.8 Puntos de Corte – P&F

MSi	MSo	CO	CMCP	WE
3.312	3.315	3.308	3.278	3.3028

Tabla 3.9 Resultados para el Modelo 1 – P&F

AVTI					ECM * 10^2				
MSi	MSo	CO	CMCP	WE	MSi	MSo	CO	CMCP	WE
0.05	0.05	0.05	0.05	0.05	4.223	4.218	4.235	4.229	4.212

Tabla 3.10 Resultados para el Modelo 2 – P&F

	R	2	5	10	14	15	18	20
AVTI	MSi	0.036	0.035	0.034	0.039	0.034	0.034	0.036
	MSo	0.037	0.035	0.038	0.034	0.032	0.036	0.036
	CO	0.035	0.037	0.037	0.066	0.081	0.184	0.120
	CMCP	0.035	0.039	0.041	0.058	0.060	0.087	0.132
	WE	0.037	0.035	0.035	0.043	0.051	0.081	0.127
OP	MSi	0.068	0.310	0.419	0.452	0.453	0.462	0.468
	MSo	0.071	0.311	0.418	0.448	0.453	0.468	0.470
	CO	0.071	0.307	0.415	0.452	0.454	0.469	0.472
	CMCP	0.071	0.306	0.416	0.445	0.456	0.462	0.477
	WE	0.070	0.309	0.416	0.450	0.455	0.469	0.471
ECM*10^2	MSi	4.837	5.010	4.676	4.679	4.674	4.577	4.588
	MSo	4.771	4.877	4.722	4.613	4.565	4.537	4.540
	CO	4.774	4.969	4.818	6.335	7.018	12.541	10.697
	CMCP	4.905	5.743	6.123	7.252	7.377	8.940	11.400
	WE	4.827	4.886	4.810	4.943	5.525	7.0134	9.491

En la Fig. 3.7, se puede observar que:

- Las medidas de desempeño de MSi y MSo sólo se ven ligeramente afectadas por *R*;
- MSi, MSo, CO y WE presentan el mismo comportamiento para *R* [2,10], pero el AVTI y ECM de CO y WE aumentan para *R*> 10 y *R* > 14, respectivamente;
- En general, la función CMCP muestra el comportamiento más pobre con respecto al ECM;
- Los valores de OP son similares para todas las técnicas analizadas;

En contraste con el caso lineal, la inicialización del problema de estimación robusta con la solución del procedimiento de RD clásico, como fue sugerido por Chen y co. (2013), aumenta el tiempo de cómputo en comparación con los requerimientos de MSo (ver Tabla 3.11).

Tabla 3.11 Promedio de Tiempos de Ejecución (s) para el Modelo 2 – P&F

MSi	MSo	CO	CMCP	WE
221.3	372.1	896.2	808.3	809.9

A continuación se presentan los resultados obtenidos con el Modelo 3. La Tabla 3.12 contiene las medidas de desempeño para diferentes valores de *K*. También los registros del AVTI y ECM se ilustran en la Fig. 3.8, y los tiempos de ejecución promedio se indican en la Tabla 3.13.

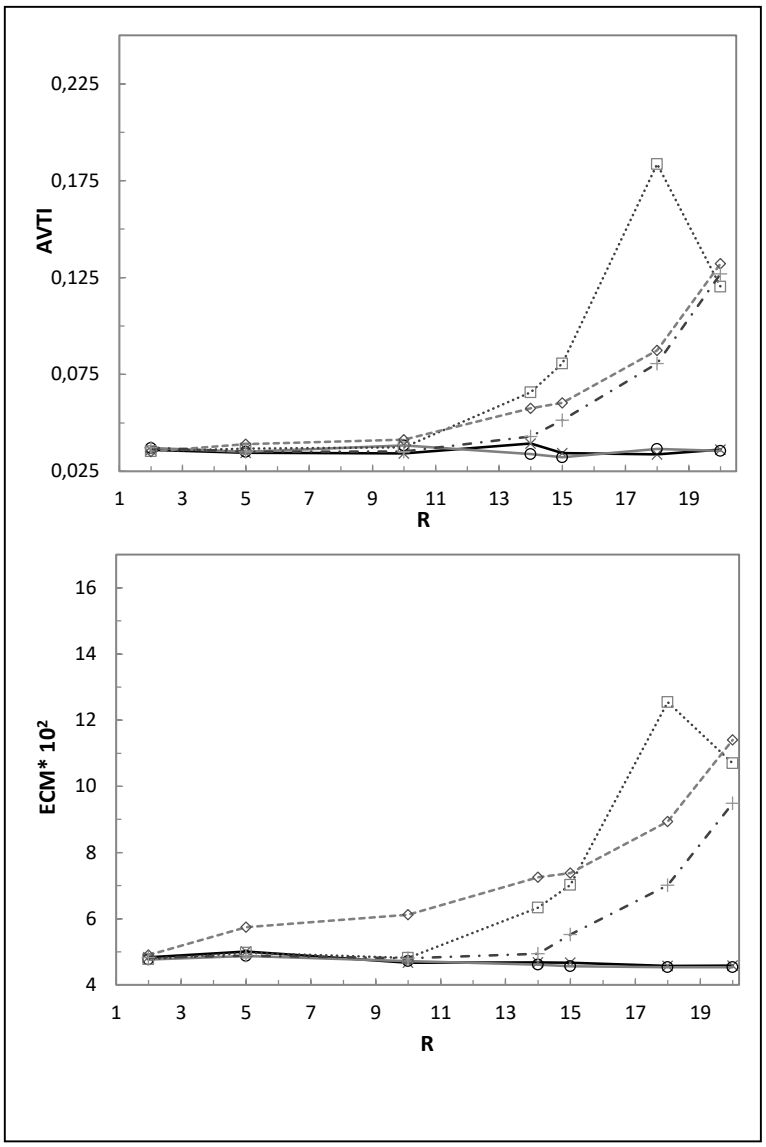


Figura 3.7 AVTI y ECM para el Modelo 2 – P&F

Para el Modelo 3, los resultados del ejemplo no lineal proporcionan las mismas conclusiones obtenidas al aplicar las metodologías al SMN. Del análisis de las medidas de desempeño, no se puede verificar una superioridad clara de una técnica sobre otra para el rango estudiado de valores de *R*. En cuanto al tiempo de cómputo, los requerimientos del procedimiento MSi son los más bajos. Además, el tiempo de ejecución de las estrategias que inicializan el problema de estimación con la solución obtenida aplicando la función CM es mayor que el tiempo consumido por MSo.

Tabla 3.12 Resultados para el Modelo 3 – P&F

	K	1	2	3	4	5	6	7	8	9	10
AVTI	SiM	0.034	0.040	0.048	0.048	0.037	0.035	0.034	0.034	0.037	0.035
	SoM	0.035	0.040	0.047	0.043	0.038	0.036	0.035	0.036	0.036	0.034
	CO	0.034	0.039	0.042	0.047	0.044	0.040	0.035	0.038	0.036	0.046
	CMCP	0.036	0.036	0.039	0.049	0.052	0.060	0.067	0.064	0.068	0.071
	WE	0.033	0.038	0.043	0.051	0.045	0.040	0.034	0.033	0.034	0.035
OP	SiM	0.000	0.000	0.074	0.953	0.999	1.000	1.000	1.000	1.000	1.000
	SoM	0.000	0.000	0.073	0.968	0.999	1.000	1.000	1.000	1.000	1.000
	CO	0.000	0.000	0.071	0.963	0.995	1.000	1.000	1.000	1.000	1.000
	CMCP	0.000	0.000	0.070	0.949	0.988	0.999	1.000	1.000	1.000	1.000
	WE	0.000	0.000	0.073	0.964	0.995	1.000	1.000	1.000	1.000	1.000
ECM* 10^2	SiM	4.392	6.638	9.188	8.395	5.087	4.625	4.665	4.575	4.646	4.671
	SoM	4.463	6.572	8.636	7.394	4.828	4.656	4.527	4.562	4.633	4.622
	CO	4.521	6.713	7.923	8.020	6.274	5.181	4.693	4.808	4.789	5.191
	CMCP	4.817	6.219	7.565	9.033	10.036	10.683	10.942	11.492	11.455	11.897
	WE	4.522	6.463	7.865	7.895	6.312	5.186	4.656	4.567	4.563	4.593

Tabla 3.13 Promedio de Tiempos de Ejecución (s) para el Modelo 3 – P&F

MSi	MSo	CO	CMCP	WE
230.6	384.3	770.1	621.3	615.1

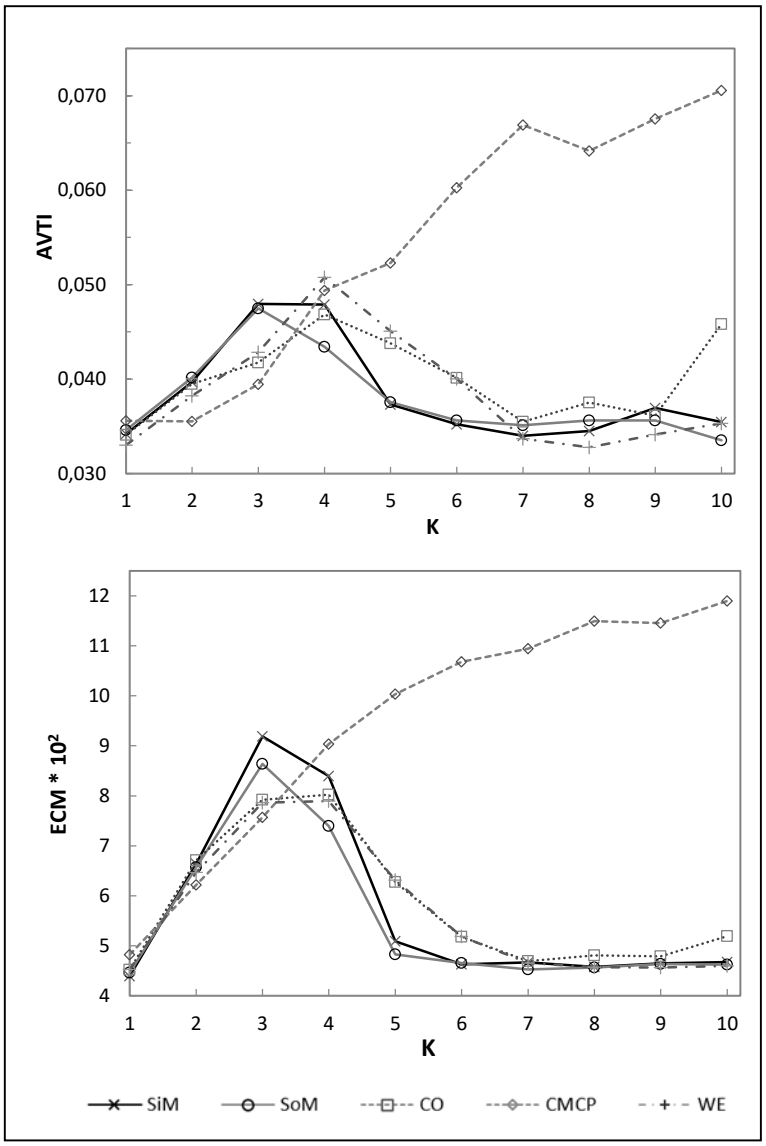


Figura 3.8. AVTI y ECM para el Modelo 3 – P&F

3.8 Conclusiones

En este capítulo se presentan dos metodologías de RDR de diferente complejidad y se compara su desempeño con los correspondientes a las estrategias que aplican los M-estimadores que han aparecido en la literatura de RDR durante la última década. Todas las estrategias se sintonizan para tener el mismo desempeño cuando las mediciones no presentan valores atípicos y se analiza su comportamiento para dos modelos de error en las mediciones.

Los resultados muestran que los M-estimadores monótonos y redescendentes se comportan de manera diferente aunque se utilice la misma inicialización del problema de optimización. En este sentido, el AVTI y ECM siempre aumentan para CMCP para valores crecientes de contaminación. En contraste, CO y WE son más robustos.

Cuando los errores de medición provienen de una contaminación normal, MSi y MSo son más robustos que las metodologías basadas en los M-estimadores WE y CO para todos los valores de contaminación probados. Si estos errores no obedecen a la distribución antes mencionada, el comportamiento de las estrategias cambia con la contaminación de manera similar, y no se puede establecer una superioridad clara de una técnica sobre las otras.

En general, el ECM más bajo se obtiene utilizando MSo, y MSi consume el menor tiempo de cómputo. Teniendo en cuenta tanto las medidas de desempeño como la carga computacional, el procedimiento MSi aparece como una alternativa eficiente para resolver el tipo de problemas bajo análisis. Éste proporciona buenas estimaciones para las mediciones reconciliadas, y su carga computacional es la más baja gracias a los beneficios de emplear la inicialización robusta del problema de estimación calculada en el Paso 1 del procedimiento.

El tratamiento simultáneo de los ESE y los errores sistemáticos que persisten en el tiempo, tales como sesgos y derivas, requieren una estrategia diferente. Ésta se presenta en el Capítulo 5 de esta tesis.

3.9 Notación

$\mathbf{a}$	Vector de ajuste de las mediciones
c	Parámetro del M-estimador
$\mathbf{f}$	Sistema de restricciones de igualdad
f_0	Función de densidad de probabilidad
$\mathbf{h}$	Sistema de restricciones de desigualdad
I	Número de variables medidas
K	Magnitud representativa del ESE Modelo 3
N	Número de réplicas de la variable medida
Ns	Número de simulaciones
R	Magnitud representativa del ESE Modelo 2
U	Número de variables no medidas
$\mathbf{u}$	Vector de variables no medidas
$\hat{\mathbf{u}}$	Vector estimado de las variables no medidas
$\mathbf{u}^{\mathbf{U}}$	Límite superior de las variables no medidas
$\mathbf{u}^{\mathbf{L}}$	Límite inferior las variables no medidas
W	Función de Peso del M-estimador
w_i	Peso de la i-ésima medición
$\mathbf{X}$	Vector de variables medidas
$\hat{\mathbf{x}}$	Vector reconciliado de las variables medidas
$\mathbf{x}^{\mathbf{U}}$	Límite superior de las variables medidas
$\mathbf{x}^{\mathbf{L}}$	Límite inferior de las variables medidas
y_{ip}	Medición de la *i*-ésima variable en el *p*-ésimo intervalo de muestreo
$\mathbf{Y}$	Vector de mediciones

ε_{ip}	Errores aleatorio de la *i*-ésima variable en el tiempo *p*
V	Varianza de la M-estimación
V_0	Varianza de la estimación obtenida con la función CM
$\sigma_{y,i}$	Desvio estándar de la medición i-ésima
ρ	Función de pérdida del M-estimador
ξ	Tasa de contaminación
ψ	Función Influencia del M-estimador
L	Función de Máxima Verosimilitud
N	Distribución normal

3.10 Acrónimos

AVTI	Número Promedio de Errores Tipo I
BW	Función Biweight
C	Función de Cauchy
CM	Cuadrados Mínimos
CMCP	Cuadrados Mínimos Cuasi Ponderado
CO	Correntropía
ECM	Error Cuadrático Medio
Ef	Eficiencia Asintótica
ESE	Error Sistemático Esporádico
FF	Fair Function
HA	Función de Hampel

HU	Función de Huber
MSi	Método Simple
MSo	Método Sofisticado
OP	Desempeño Global
P&F	Ejemplo de Pai and Fisher
RD	Reconciliación de Datos
RDR	Reconciliación de Datos Robusta
SMN	Red de Ingreso de Vapor
WE	Función de Welsch

Capítulo 4

Test Robusto de las Mediciones

4 Test Robusto de las Mediciones

4.1 Introducción

En este capítulo se presenta el Test Robusto de las Mediciones (TRM). Su propósito es detectar e identificar valores atípicos en las mediciones con alta eficiencia, aún en presencia de múltiples Errores Sistemáticos Esporádicos (ESE).

En el marco de la Reconciliación de Datos Clásica (RDC), las técnicas para el tratamiento de ESE tienen como objetivo reconocer las mediciones atípicas presentes en el vector actual de observaciones y evitar su uso en el procedimiento de estimación que minimiza la función Cuadrados Mínimos (CM).

Los métodos de tratamiento de ESE existentes sólo pueden analizar la posible presencia de errores en las observaciones categorizadas como redundantes porque requieren que existan al menos dos caminos alternativos para estimar el valor de una variable medida. Es decir, es necesario disponer del valor de la medición y además, del valor calculado empleando ecuaciones que comprendan exclusivamente variables medidas redundantes. La precisión de estas variables mejora al ejecutarse el procedimiento de RDC, por lo tanto resulta importante incrementar la Redundancia Espacial (RE) en la información del proceso mediante la instalación de instrumentos en variables no medidas o la colocación de sensores por duplicado (Romagnoli y Sánchez, 2000).

La baja redundancia de las variables medidas redundantes afecta el desempeño de los métodos previamente citados (Narasimhan y Jordache, 2000). Maronna y Arcas (2009) desarrollaron una metodología sencilla para el cálculo de la redundancia en

sistemas lineales. Estos autores también demostraron que existe una relación directa entre la redundancia y la probabilidad de detección de mediciones atípicas para sistemas lineales.

Generalmente los métodos empleados para analizar la presencia de ESE se basan en la aplicación de test de hipótesis estadísticos considerando un modelo del proceso que comprende sólo variables medidas. Cuando los modelos están conformados por restricciones lineales, los estadísticos de estos test siguen una distribución de probabilidad conocida. Si ésta no es la situación de partida, las ecuaciones se linealizan y además se eliminan las variables no medidas.

Algunos de los test más utilizados para la detección de mediciones atípicas en el marco de la RDC son: el Test Global, Test Nodal, Test de las Mediciones (TM), Test Razón de Máxima Verosimilitud (TRMV), Test de las Componente Principales, etc. Estos test pueden formar parte de estrategias iterativas de detección e identificación. Si bien existen diversas técnicas para localizar los errores sistemáticos, muy pocas son capaces de realizar la detección e identificación conjuntas con un elevado grado de acierto.

El TM (Mah y Tamhane, 1982) se destaca por tener la capacidad de detectar e identificar simultáneamente cuáles son las variables medidas redundantes, contenidas en el vector de observaciones actual, que presentan mediciones atípicas. Para su formulación se necesita disponer del vector de estimaciones, $\hat{\mathbf{x}}$, obtenido al efectuar la RDC con la función CM. La presencia de observaciones con errores sistemáticos deteriora la exactitud del $\hat{\mathbf{x}}$ como consecuencia de la dispersión del error en las estimaciones, lo que origina una reducción en la eficiencia del TM.

Se han propuesto diferentes variantes del TM con el objetivo de mejorar su comportamiento, destacándose entre estas metodologías el Test de las Mediciones Iterativo Modificado (TMIM) (Serth y Heenan, 1986). Éste se basa en eliminar secuencialmente la variable medida más sospechosa del conjunto de ecuaciones de reconciliación. No obstante Crowe (1988) demostró que este tipo de eliminación puede no conducir a la identificación correcta de las mediciones atípicas. Por su parte, Özyurt y Pike (2004) compararon el comportamiento del TMIM con otra metodología que emplea una regla de rechazo basada en conceptos de Estadística Robusta y obtuvieron resultados similares pero empleando menor tiempo de cómputo.

Se sabe que la Estadística Robusta proporciona estimaciones más exactas y precisas aún en presencia de observaciones atípicas, por ello en esta tesis se formula un nuevo test estadístico, denominado Test Robusto de las Mediciones (TRM), con los siguientes objetivos: que sea capaz de detectar un alto porcentaje de ESE con un bajo grado de falsas alarmas, en presencia de múltiples ESE, aun cuando la redundancia de las variables medidas sea baja. Ésta se cuantifica siguiendo el desarrollo presentado por Maronna y Arcas (2009) para sistemas lineales, y se lo extiende en la presente tesis para abordar sistemas no lineales.

Además, en este capítulo se analiza de forma exhaustiva el comportamiento del TRM para procesos que comprenden variables con distinta RE y cuya operación se representa mediante sistemas de ecuaciones algebraicas lineales y no lineales. También se aborda el desempeño del nuevo test en sistemas para los cuales las estrategias existentes evidenciaron problemas de identificación, tales como los procesos con variables equivalentes o corrientes paralelas.

4.2 Test de las Mediciones Clásico

El TM fue desarrollado por Mah y Tamhane (1982) para detectar la presencia de mediciones atípicas en un vector de observaciones. Las hipótesis estadísticas del test son las siguientes:

H_0: No hay mediciones atípicas en el vector de observación actual

H_1: Hay mediciones atípicas en el vector de observación actual

En los subsiguientes desarrollos consideraremos, por simplicidad, un modelo de proceso lineal representado por la matriz $\mathbf{A_1}$, que relaciona *I* variables medidas, las cuales siguen una distribución $\mathcal{N}(\mathbf{x}, \mathbf{\Sigma})$ en ausencia de errores sistemáticos, siendo $\mathbf{\Sigma}$ la matriz de covarianza de las mediciones considerada diagonal y σ_y el desvío estándar de las observaciones.

Para el cálculo del estadístico se propusieron los siguientes pasos:

1) Cálculo del valor estimado de las variables medidas

Para un instante dado se resuelve el siguiente problema de RDC utilizando como estimador la función CM.

$$\hat{\mathbf{x}}^{CM} = \underset{\mathbf{x}}{Min} \quad \sum_{i=1}^{I} \left(\frac{y_i - x_i}{\sigma_{y,i}} \right)^2, \quad s.t. \quad \mathbf{A}_1\mathbf{x} = \mathbf{0} \tag{4.1}$$

2) Cálculo del ajuste y de la covarianza del ajuste

$$\mathbf{a} = \mathbf{y} - \hat{\mathbf{x}}^{CM} \tag{4.2}$$

Reemplazando $\hat{\mathbf{x}}^{CM}$ por la solución analítica obtenida usando el método de los multiplicadores de Lagrange se obtiene la siguiente expresión del ajuste:

$$\mathbf{a} = \mathbf{y} - \left[\mathbf{y} - \boldsymbol{\Sigma}\mathbf{A}_1^T(\mathbf{A}_1\boldsymbol{\Sigma}\mathbf{A}_1^T)^{-1}\mathbf{A}_1\mathbf{y}\right]$$

$$\mathbf{a} = \boldsymbol{\Sigma}\mathbf{A}_1^T\mathbf{V}^{-1}\mathbf{A}_1\mathbf{y} \tag{4.3}$$

La matriz de covarianza del ajuste, $\mathbf{Q}$, resulta ser:

$$\mathbf{Q} = \boldsymbol{\Sigma}\mathbf{A}_1^T\mathbf{V}^{-1}\mathbf{A}_1\mathrm{Cov}(\mathbf{y})\left(\boldsymbol{\Sigma}\mathbf{A}_1^T\mathbf{V}^{-1}\mathbf{A}_1\right)^T$$

$$\mathbf{Q} = \boldsymbol{\Sigma}\mathbf{A}_1^T\mathbf{V}^{-1}\mathbf{A}_1\boldsymbol{\Sigma}^T \tag{4.4}$$

donde $\mathrm{Cov}(\bullet)$ representa la covarianza del argumento. De la Ec. 4.3 se observa que el vector de ajustes no es más que una transformación lineal del vector de mediciones, por lo tanto seguirá la misma distribución de probabilidad que las mediciones es decir: $\mathbf{a} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q})$.

3) Formulación del estadístico

El estadístico de la *i*-ésima variable se calcula como:

$$\tau_{a,i} = \frac{|a_i|}{\sqrt{Q_{ii}}} \quad i = 1, 2...I \tag{4.5}$$

Si se satisface H_0, $\tau_{a,i} \sim \mathcal{N}(0,1)$. Su valor se compara con el crítico, $t_{c,1-\beta/2}$, donde β se obtiene reemplazando m por I en la Ec. 2.12.

La utilización de este test resulta atractiva, pues permite localizar mediciones atípicas sin necesidad de recurrir a procedimientos iterativos. Sin embargo, la presencia de un solo ESE o de múltiples errores perjudica la exactitud de $\hat{\mathbf{x}}^{CM}$, y el test produce un alto porcentaje de falsas alarmas, lo cual hace inviable su empleo (Serth y Heenan, 1986). Con el objetivo de mejorar el comportamiento del TM se han propuesto estrategias iterativas que lo utilizan, no obstante sus resultados no son satisfactorios (Crowe 1996; Narasimhan y Jordache, 2000).

4.3 Test de las Mediciones para Ventanas de Datos

Con el fin de aprovechar la redundancia temporal (RT) de las observaciones del proceso, se propone en esta tesis el Test de las Mediciones en una Ventana de Datos (TMV). Las hipótesis estadísticas del test son las siguientes:

H_0: No hay mediciones atípicas en el vector de observación actual de la Ventana

H_1: Hay mediciones atípicas en el vector de observación actual de la Ventana

Para el intervalo de muestreo *j*, se denomina Y_{ob} a la matriz de dimensión (*I* x *N*) que contiene los últimos *N* vectores de medición obtenidos hasta *j* inclusive, $\mathbf{y_p}$ {$p = j-N+1 \ldots j$}. Por otra parte, $\overline{\mathbf{y}}_j$ representa el vector promedio de las mediciones de dicha ventana, es decir:

$$\overline{\mathbf{y}}_j = \frac{\sum_{p=j-N+1}^{j} \mathbf{y_p}}{N} \tag{4.6}$$

Esta nueva variable aleatoria seguirá una distribución $\mathcal{N}(\mathbf{x}, \frac{\mathbf{\Sigma}}{\mathcal{N}})$.

Al igual que en el TM, el estadístico se obtiene realizando los siguientes tres pasos:

1) Cálculo del valor estimado de las variables medidas

El vector de estimaciones de las variables medidas en el intervalo de tiempo *j*, $\overline{\mathbf{x}}_j^{CM}$, es la solución del Problema 4.7, cuando se utiliza como estimador la función CM:

$$\begin{aligned} \overline{\mathbf{x}}_j^{CM} = & \underset{\mathbf{x_j}}{Min} \quad \sum_{i=1}^{I} \left(\frac{\overline{y}_i - x_i}{\sigma_{y,i}} \right)^2 \\ & s.t. \\ & \mathbf{A}_1 \mathbf{x} = 0 \end{aligned} \tag{4.7}$$

2) Cálculo del ajuste y de la covarianza del ajuste

El vector de ajuste de las mediciones, $\bar{\mathbf{a}}_j^{CM}$, se define como:

$$\begin{aligned}\bar{\mathbf{a}}_j^{CM} &= \mathbf{y}_j - \bar{\mathbf{x}}_j^{CM}\mathbf{y}_j - (\bar{\mathbf{y}}_j - \boldsymbol{\Sigma}\mathbf{A}_1^T\mathbf{V}^{-1}\mathbf{A}_1\bar{\mathbf{y}}_\mathbf{j}) \\ \bar{\mathbf{a}}_j^{CM} &= \mathbf{y}_j - (\mathbf{I} - \boldsymbol{\Sigma}\mathbf{A}_1^T\mathbf{V}^{-1}\mathbf{A}_1)\bar{\mathbf{y}}_j = \mathbf{y}_j - \mathbf{Z}\,\bar{\mathbf{y}}_j\end{aligned} \tag{4.8}$$

La matriz de covarianza del ajuste, $\bar{\mathbf{Q}}$, resultante es (Llanos y co., 2017):

$$\begin{aligned}\bar{\mathbf{Q}} &= \mathrm{Cov}(\bar{\mathbf{a}}_j^{CM}) = \mathrm{Cov}\left[\mathbf{y}_j - \mathbf{Z}\,\bar{\mathbf{y}}_j\right] \\ \bar{\mathbf{Q}} &= \mathrm{Cov}\left[\mathbf{y}_j - \mathbf{Z}\,\frac{\sum_{p=j-N+1}^{p}\mathbf{y}_p}{N}\right] \\ \bar{\mathbf{Q}} &= \mathrm{Cov}\left[\mathbf{y}_j - \mathbf{Z}\left(\frac{\sum_{p=j-N}^{p}\mathbf{y}_p + \mathbf{y}_j}{N}\right)\right] \\ \bar{\mathbf{Q}} &= \mathrm{Cov}\left[\left(\mathbf{I} - \frac{\mathbf{Z}}{N}\right)\mathbf{y}_j - \mathbf{Z}\left(\frac{\sum_{p=j-N}^{p}\mathbf{y}_p}{N}\right)\right] \\ \bar{\mathbf{Q}} &= \mathrm{Cov}\left[\left(\mathbf{I} - \frac{\mathbf{Z}}{N}\right)\mathbf{y}_j\right] + \mathbf{Cov}\left[\frac{\mathbf{Z}}{N}\left(\sum_{p=j-N}^{p}\mathbf{y}_p\right)\right] \\ \bar{\mathbf{Q}} &= \left(\mathbf{I} - \frac{\mathbf{Z}}{N}\right)\mathrm{Cov}(\mathbf{y}_j)\left(\mathbf{I} - \frac{\mathbf{Z}}{N}\right)^T + \frac{\mathbf{Z}}{N}\mathrm{Cov}\left(\sum_{p=j-N}^{p}\mathbf{y}_p\right)\left(\frac{\mathbf{Z}}{N}\right)^T \\ \bar{\mathbf{Q}} &= \left(\mathbf{I} - \frac{\mathbf{Z}}{N}\right)\Sigma\left(\mathbf{I} - \frac{\mathbf{Z}}{N}\right)^T + \frac{\mathbf{Z}}{N}\left[(N-1)\Sigma\right]\left(\frac{\mathbf{Z}}{N}\right)^T\end{aligned} \tag{4.9}$$

Reordenando:

$$\bar{\mathbf{Q}} = \left[\frac{N\mathbf{I} - \mathbf{Z}}{N}\right]\Sigma\left[\frac{N\mathbf{I} - \mathbf{Z}}{N}\right]^T + \left[\frac{(N-1)}{N^2}\right]\mathbf{Z}\Sigma\mathbf{Z}^T \tag{4.10}$$

En presencia de errores aleatorios en las mediciones, $\bar{\mathbf{a}}_j^{CM} \sim \mathcal{N}(\mathbf{0}, \bar{\mathbf{Q}})$

3) Formulación del estadístico

A partir del resultado anterior, se define el siguiente estadístico univariado $\bar{\tau}_{a,i}$ para testear la *i*-ésima observación en el *j*-ésimo intervalo de tiempo

$$\bar{\tau}_{a,i} = \frac{\left|\bar{a}_{ij}^{CM}\right|}{\sqrt{\bar{Q}_{ii}}} \quad i = 1,2...I \tag{4.11}$$

Si se satisface H_0, $\bar{\tau}_{a,i} \sim \mathcal{N}(0,1)$. Su valor se compara con el crítico $t_{c,1-\beta/2}$, donde β se obtiene reemplazando m por I en la Ec. 2.12, y se declara la presencia de un error sistemático en la *i*-ésima observación si el valor del estadístico es mayor al valor umbral, $t_{c,1-\beta/2}$.

A diferencia del TM Clásico, la formulación del ajuste utilizada en el TMV permite detectar mediciones con ESE en variables no redundantes. Sin embargo, se sabe que el punto de quiebre del estimador CM es próximo a cero, lo cual implica que la sola presencia de un error sistemático invalida los supuestos bajos los cuales se desarrolla la metodología. Esto origina que la RDC produzca estimaciones sesgadas de las variables medidas que enmascaran la identificación de las observaciones que presentan ESE. Para tomar mejores decisiones respecto de la ocurrencia de este tipo de error, se propone el desarrollo de un test basado en conceptos de la Estadística Robusta.

4.4 Test Robusto de las Mediciones

Las hipótesis estadísticas del test son las siguientes:

H_0: No hay mediciones atípicas en la observación actual de la i-ésima variable

H_1: Hay mediciones atípicas en la observación actual de la i-ésima variable

Comprende la ejecución de los pasos que se describen a continuación.

1) Cálculo del valor estimado de las variables medidas

Al igual que en los dos test anteriores, en este procedimiento es necesario inicialmente calcular el valor estimado de las observaciones. Debido a las dificultades expuestas sobre el uso de la función CM, se propone calcular una estimación robusta de las mediciones que satisfaga el modelo del proceso. Además, se considerarán las observaciones contenidas en una ventana de datos de tamaño *N*, debido a las ventajas que se consiguen con el empleo de la RT, que fueron expuestas en la Sección 4.3.

En el Capítulo 3 de la tesis se desarrollaron y compararon diferentes estrategias de RDR. El estudio comparativo de desempeño consideró índices relacionados con la precisión y exactitud de las estimaciones y el tiempo de cómputo. El análisis puso en evidencia las cualidades de la metodología MSi para ser utilizada en procedimientos de optimización en línea. Por tal motivo se la selecciona para calcular las estimaciones del estado del proceso en el *j*-ésimo intervalo de muestreo. Se denomina $\hat{x}_j^{SIM}$ al vector de mediciones reconciliadas en el intervalo de tiempo *j*, que se obtiene mediante la resolución secuencial de los siguientes problemas de optimización

$$\tilde{y}_i = Min \quad \sum_{p=j-N+1}^{j} \rho_{BW}\left(\frac{y_{ip}-\tilde{y}_i}{\sigma_{y,i}}\right) \tag{4.12}$$

$$[\hat{\mathbf{x}}_j^R, \hat{\mathbf{u}}_j^R] = \underset{x_j,u_j}{Min} \quad \sum_{i=1}^{I} \rho_{HU}\left(\frac{\tilde{y}_i - x_i}{\sigma_{y,i}}\right)$$

$$st. \tag{4.13}$$

$$\mathbf{A}_1\mathbf{x}=\mathbf{0}$$

donde $\tilde{y}_i$ representa la mediana robusta de la ventana de observaciones.

2) Cálculo del ajuste robusto y de la covarianza del ajuste

El ajuste robusto de las mediciones, $\mathbf{a}_j^R$ se define a continuación:

$$\mathbf{a}_j^R = \mathbf{y}_j - \hat{\mathbf{x}}_j^{SIM} \tag{4.14}$$

y su matriz de covarianza se denota $\mathbf{Q}^R$. Si las observaciones presentan errores aleatorios de distribución normal y la operación del proceso puede representarse mediante un conjunto de ecuaciones algebraicas lineales, el vector $\mathbf{a}_j^R$ sigue aproximadamente una distribución $\mathcal{N}(\mathbf{0},\ \mathbf{Q}^R)$. Esto ocurre porque $\hat{\mathbf{x}}_j^{SIM}$ se estima usando una transformación lineal de la mediana robusta, la cual tiende asintóticamente a la distribución normal. Esta suposición es válida para cualquier M-estimador utilizado en un problema de RDR y ha sido rigurosamente demostrada (Maronna y co., 2006).

La matriz $\mathbf{Q}^R$ es desconocida pero puede calcularse una estimación robusta de la misma en el tiempo *j*, que se denota $\hat{\mathbf{Q}}_j^R$. Para ello, se guarda una matriz de ajuste $\mathbf{A}_j^R$ que está formada por los últimos $\mathbf{a}_p^R$ ($p = j\text{-}N+1,\ldots, j$) vectores,

$$\mathbf{A}_j^R = [\mathbf{a}_{j-N+1}^R, \mathbf{a}_{j-N+2}^R, \ldots \mathbf{a}_j^R] \tag{4.15}$$

La mediana normalizada de las desviaciones absolutas alrededor de la mediana para el ajuste de la *i*-ésima variable, *MADN* ($\mathbf{a}_i^R$), se estima usando la *i*-ésima fila de $\mathbf{A}_j^R$ de la siguiente forma:

$$MAD\left(\mathbf{a}_i^R\right) = Med\left[\left|\mathbf{A}_j^R(i,:) - Med\left(\mathbf{A}_j^R(i,:)\right)\right|\right] \tag{4.16}$$

$$MADN\left(\mathbf{a}_i^R\right) = \frac{MAD\left(\mathbf{a}_i^R\right)}{0.675} \tag{4.17}$$

y el cuadrado de la *MADN* ($\mathbf{a}_i^R$) (i=1,…,I) se emplea para obtener una estimación del vector de escala $\hat{\sigma}_a^2$. A continuación, se evalúa la matriz $\hat{\mathbf{Q}}_j^R$ como sigue:

$$\hat{\mathbf{Q}}_j^R = \hat{\boldsymbol{\sigma}}_a^2 \left\{ \frac{\text{ave}\left[\psi(\mathbf{A}_j^R)/\hat{\boldsymbol{\sigma}}_a\right]^2}{\left(\text{ave}\left[\psi'(\mathbf{A}_j^R/\hat{\boldsymbol{\sigma}}_a)\right]\right)^2} \right\}^T \tag{4.18}$$

donde ψ es la derivada de la función BW (Ec 3.26) y ave(•) representa el promedio de la muestra para el argumento. En presencia de errores aleatorios esta varianza tiende a la varianza muestral obtenida a partir de la Estadística Clásica, por lo tanto tiende a una distribución chi-cuadrado, es decir, $\hat{Q}_j^R \sim \chi^2$.

3) Formulación del estadístico

El estadístico del TRM, $\tau_{i,j}^R$, se formula usando $\hat{\mathbf{Q}}_j^R$ de la siguiente manera:

$$\hat{\tau}_{i,j}^R = \frac{\left|a_{ij}^R\right|}{\sqrt{\hat{Q}_j^R\big|_{ii}}} \tag{4.19}$$

Este $\tau_{i,j}^R$ sigue la distribución de Student con un número de grados de libertad *df = N-1*, es decir:

$$\hat{\tau}_{i,j}^R \sim t_{N-1} \tag{4.20}$$

La distribución del $\tau_{i,j}^R$ se verifica si $\mathbf{a}_j^R$ sigue asintóticamente una distribución normal y $\hat{Q}_j^R \sim \chi^2$. Esto se demuestra en el Anexo 2. El valor del estadístico se compara con el estadístico crítico τ_c^R para un nivel de significancia del test α, es decir, con $t_{N-1,1-\alpha/2}$.

En los sistemas de ecuaciones no lineales la distribución de a_j^R no se conoce, se recomienda el siguiente procedimiento para estimar el valor crítico. Inicialmente, se asume que a^R sigue asintóticamente la distribución normal. Esta hipótesis de trabajo permite calcular $t_{N-1,1-\alpha/2}$. Luego esta suposición debe ser validada usando un conjunto de muestras de a^R. Si la hipótesis de trabajo se rechaza, por ejemplo, la probabilidad

experimental de $\{|t_i^R| > t_{N-1,1-\alpha/2}\}$ es diferente de α, la muestra se usa para calcular el valor del estadístico crítico. Las técnicas de estimación de la función de densidad de probabilidad tipo kernel pueden aplicarse con este propósito.

4.5 Cuantificación de la Redundancia Espacial de las Variables Medidas

Para un dado proceso, la RE de las variables medidas depende de la cantidad de sensores instalados y de la precisión de los mismos. Cuanto mayor es la RE de una variable, mayor será la precisión de su estimación, o sea el desvío estándar de la estimación será notablemente menor que el de la medición. Madron (1992) definió una medida práctica de la RE, denominada Ajustabilidad de la i-ésima Variable Medida $(Ajus_i)$, de la siguiente manera:

$$Ajus_i = \left(1 - \frac{\sigma_{\hat{x},i}}{\sigma_{y,i}}\right) > Ajus_c, \tag{4.21}$$

donde $\sigma_{\hat{x},i}$ representa el desvío de la estimación, $\sigma_{y,i}$ el desvío de la medición y $Ajus_c$ un valor umbral que puede variar en el intervalo (0,1). La comparación de este último coeficiente con $Ajus_i$ permite concluir respecto de la RE de la variable. Por ejemplo, si se considera $Ajus_c = 0{,}1$, un $Ajus_i$ inferior a ese valor umbral implica que el ajuste realizado a la variable es insignificante y su RE prácticamente nula.

Charpentier y co. (1991) presentaron una relación que permite identificar las mediciones con baja redundancia. Definieron la Detectabilidad de Error en la *i*-ésima Variable Medida (Det_i) mediante la siguiente ecuación:

$$Det_i = \sqrt{\left(1 - \frac{\sigma^2_{\hat{x},i}}{\sigma^2_{y,i}}\right)} \tag{4.22}$$

Debido a que la inconsistencia de una restricción puede relacionarse a la presencia de un error sistemático, la posibilidad de detectarlo depende de su contribución al incumplimiento de dicha restricción. A medida que Det_i es mayor, el error sistemático se detectada con mayor facilidad o se pueden detectar errores no aleatorios de menor magnitud.

Por su parte, Maronna y Arcas (2009) presentaron una metodología de regresión múltiple que permite evaluar la RE en sistemas lineales. En este trabajo se define la Redundancia Espacial de la *i*-ésima Variable, RE_i, como:

$$RE_i = 1 - \frac{\sigma^2_{\hat{x},i}}{\sigma^2_{y,i}} \tag{4.23}$$

El desarrollo presentado por estos autores permite evaluar la Ec. 4.23 en función de la matriz $\mathbf{A_1}$, que representa el modelo del proceso, y $\mathbf{\Sigma}$.

Narasimhan y Jordache (2000) indican que en sistemas no lineales la redundancia o una medida de la redundancia sólo pueden evaluarse después de resolver el problema de RDC para un conjunto representativo de mediciones. Esto se debe a que, a diferencia de los sistemas lineales, no se tiene una expresión analítica para la matriz de covarianza de las estimaciones. A continuación, se repasa el desarrollo de Maronna y Arcas (2009) en el cual se obtiene una expresión para el cálculo de la RE en sistemas lineales y, en base a este procedimiento, se extiende la expresión a sistemas no lineales.

4.5.1 Redundancia Espacial de las Variables Medidas en Sistemas Lineales

Dado el modelo del proceso representado por el siguiente conjunto de *M* ecuaciones lineales

$$\mathbf{A}\mathbf{v} = \mathbf{0}, \tag{4.24}$$

éste puede reformularse de la siguiente manera:

$$\mathbf{A_1 x} + \mathbf{A_2 u} = \mathbf{0}, \tag{4.25}$$

siendo **x** (*I*x1) y **u** (*U*x1) los vectores correspondientes a las variables medidas y no medidas, respectivamente, y $\mathbf{A_1}$ y $\mathbf{A_2}$ matrices de dimensión compatible.

Por definición de subespacio nulo de una matriz se sabe que:

Definición 4.1: Subespacio nulo

Dada una matriz $\mathbf{A} \in R^{M\times(I+U)}$ $\exists$ **v** /

$$Null(\mathbf{A}) = Ker(\mathbf{A}) = \left\{ \mathbf{v} \in \Re^{I+U} \,\middle|\, \mathbf{Av} = \mathbf{0}_{\Re^M} \right\}.$$

Si denominamos $q=I+U-M$ y $\mathbf{b_1,\ldots,b_q}$ a los vectores base del subespacio, entonces cualquier vector **v** que satisfaga la definición anterior puede ser expresado como una combinación lineal de estos vectores:

$$\mathbf{v} = \sum_{j=1}^{q} \beta_j \mathbf{b}_j \quad \beta_j \in R \tag{4.26}$$

Llamando **B** a la matriz formada por los *q* vectores $\mathbf{b}_j$ y **β** al vector que contiene los β_j, la Ec 4.26 puede reformularse de manera matricial como:

$$\mathbf{v} = \mathbf{B}\boldsymbol{\beta} \tag{4.27}$$

Si dividimos a la matriz **B** en $\mathbf{B}_x$ y $\mathbf{B}_u$, donde la primera tiene dimensión $I\times q$ y la segunda $U\times q$, **B** puede reescribirse como:

$$\mathbf{B} = \left[\mathbf{b}_1, \ldots, \mathbf{b}_q \right] = \begin{bmatrix} \mathbf{B}_x \\ \mathbf{B}_u \end{bmatrix} \tag{4.28}$$

y **v** subdividirse en:

$$\mathbf{x} = \mathbf{B}_x\boldsymbol{\beta}$$

$$\mathbf{u} = \mathbf{B}_u\boldsymbol{\beta} \tag{4.29}$$

El vector de mediciones $\mathbf{y}$ del modelo de regresión lineal con la matriz predictora $\mathbf{B}_x$ resulta entonces igual a:

$$\mathbf{y} = \mathbf{B}_x\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4.30}$$

Si se denomina $\boldsymbol{\Sigma}^{1/2}$ a la matriz que contiene la raíz cuadrada de los elementos de la diagonal de la matriz Σ, es decir los $\sigma_{y,i}$, es posible estandarizar **y** como:

$$\mathbf{y}_s = \left(\boldsymbol{\Sigma}^{1/2}\right)^{-1}\mathbf{y}, \tag{4.31}$$

y reescribir la Ec. 4.31 como:

$$\mathbf{y}_s = \left(\Sigma^{1/2}\right)^{-1}\mathbf{B}_x\boldsymbol{\beta} + \left(\Sigma^{1/2}\right)^{-1}\boldsymbol{\varepsilon} \tag{4.32}$$

Si se denota $\mathbf{C} = \left(\boldsymbol{\Sigma}^{1/2}\right)^{-1}\mathbf{B}_x$ y $\boldsymbol{\varepsilon}_s = \left(\Sigma^{1/2}\right)^{-1}\boldsymbol{\varepsilon}$, y se reformula la Ec. 4.31

$$\mathbf{y}_s = \mathbf{C}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_s, \tag{4.33}$$

se estima $\boldsymbol{\beta}$ resolviendo el siguiente problema de optimización usando la función CM:

$$\left\|\mathbf{y}_s - \mathbf{C}\hat{\boldsymbol{\beta}}\right\| = Min \tag{4.34}$$

La solución del problema (4.34) es:

$$\hat{\boldsymbol{\beta}} = \mathbf{C}^{+}\mathbf{y}_s \tag{4.35}$$

donde $\mathbf{C}^{+}$ denota la pseudo inversa de $\mathbf{C}$; reemplazando $\hat{\boldsymbol{\beta}}$ en la Ec. (4.34) resulta:

$$\hat{\mathbf{y}}_s = \mathbf{C}\hat{\boldsymbol{\beta}} = \mathbf{C}\mathbf{C}^{+}\mathbf{y}_s = \mathbf{H}\mathbf{y}_s \tag{4.36}$$

siendo **H** una matriz idempotente conocida en Análisis de Regresión como matriz sombrero, que permite obtener el vector estimado como una transformación lineal de las mediciones. La varianza del vector $\hat{\mathbf{y}}_s$ resulta:

$$\mathrm{Var}(\hat{\mathbf{y}}_s) = \mathrm{Var}\left(\mathbf{H}\mathbf{y}_s\right) = \mathbf{H}\,\mathrm{Var}\left(\mathbf{y}_s\right)\mathbf{H}^T = \mathbf{H}\,\mathbf{I}\,\mathbf{H}^T = \mathbf{H} \tag{4.37}$$

Si se denomina h_i al *i*-ésimo elemento ubicado en la diagonal de **H**, la Ec. 4.23 puede reescribirse como:

$$\mathrm{RE_i} = 1 - h_i \tag{4.38}$$

Esta expresión permite calcular la $\mathrm{RE_i}$ a partir de la matriz **H**, siendo ésta función del modelo del proceso y de la matriz de covarianza de las mediciones. A continuación, se demostrará que está expresión puede extenderse a sistemas no lineales.

4.5.2 Redundancia Espacial de las Variables Medidas en Sistemas No Lineales

La linealización de un conjunto de *M* restricciones no lineales representadas por $\varphi(\mathbf{x},\mathbf{u}) = \mathbf{0}$ puede realizarse utilizando un polinomio de Taylor entorno a una estimación inicial (x_0,u_0)

$$P_{1,(x_0,u_0)}\left[\varphi(\mathbf{x},\mathbf{u})\right] = \varphi(\mathbf{x}_0,\mathbf{u}_0) + \nabla\varphi(\mathbf{x})\big|_{x_0,u_0}(\mathbf{x}-\mathbf{x}_0) + \nabla\varphi(\mathbf{u})\big|_{x_0,u_0}(\mathbf{u}-\mathbf{u}_0) \approx \mathbf{0} \tag{4.39}$$

Si se denomina $\mathbf{J}_1 = \nabla\varphi(\mathbf{x})\big|_{x_0,u_0}$ y $\mathbf{J}_2 = \nabla\varphi(\mathbf{u})\big|_{x_0,u_0}$, la Ec. 4.39 resulta igual a:

$$\varphi(\mathbf{x}_0,\mathbf{u}_0) + \mathbf{J}_1(\mathbf{x}-\mathbf{x}_0) + \mathbf{J}_2(\mathbf{x}-\mathbf{x}_0) = \mathbf{0} \tag{4.40}$$

Reordenando:

$$\mathbf{J}_1\mathbf{x} + \mathbf{J}_2\mathbf{u} = \mathbf{J}_1\mathbf{x}_0 + \mathbf{J}_2\mathbf{u}_0 - \varphi(\mathbf{x}_0,\mathbf{u}_0) = \mathbf{c} \tag{4.41}$$

Si se considera que: $\mathbf{J} = \begin{bmatrix} \mathbf{J}_1 & \mathbf{J}_2 \end{bmatrix} \in R^{M\times(I+U)}$ y $\mathbf{v}_1 = \begin{bmatrix} \mathbf{x} \\ \mathbf{u} \end{bmatrix} \in R^{(I+U)}$ el sistema de ecuaciones anterior puede escribirse como:

$$\mathbf{J}\mathbf{v}_1 = \mathbf{c} \tag{4.42}$$

A diferencia de la Ec. 4.24, la Ec. 4.42 representa un sistema lineal no homogéneo. Aplicando conceptos del Algebra Lineal,

$$\mathbf{v}_1 = \mathbf{v} + \mathbf{p}, \tag{4.43}$$

donde **v** es al subespacio nulo de **J,** mientras que **p** es un vector que particulariza al sistema; reemplazando está expresión en la Ec (4.42) resulta:

$$\mathbf{J}\left(\mathbf{v} + \mathbf{p}\right) = \mathbf{c} \tag{4.44}$$

y dado que $\mathbf{J}\mathbf{v} = \mathbf{0}$ entonces:

$$\mathbf{J}\mathbf{p} = \mathbf{c} \tag{4.45}$$

Esto implica que el subespacio nulo de la matriz **J** es el mismo independientemente de si el sistema es homogéneo o no. En consecuencia, es posible aplicar al sistema linealizado el procedimiento de Maronna y Arcas (2009) y calcular valores aproximados de RE_i.

La comprobación de la metodología propuesta se realiza utilizando como caso de estudio el ejemplo presentado por Pai and Fisher (1988), P&F. Este sistema está formado por 6 ecuaciones no lineales y 8 variables, de las cuales 5 son medidas. Se realizan 10000 simulaciones de la metodología MSi empleando mediciones típicas del sistema para N=40 y se calcula la métrica Error Cuadrático Medio de la i-ésima variable, ECM_i:

$$\mathrm{ECM_i} = \frac{1}{Ns}\sum_{k=1}^{N_s}\left(\frac{\hat{x}_{ik}^{R} - x_i}{\sigma_{y,i}}\right)^2, \tag{4.46}$$

siendo *Ns* la cantidad de simulaciones. Por otra parte, la matriz **J** del sistema ingresa al procedimiento de cálculo de Maronna y Arcas (2009), y se calculan las RE_i (i = 1…5). Los resultados de RE_i y ECM_i se presentan en la Tabla 4.1.

Tabla 4.1 Redundancia y Error Cuadrático Medio de las Variables – P&F

I	1	2	3	4	5
RE_i	0.318	0.691	0.988	0.439	0.564
ECM_i x 100	16.481	0.7911	0.0293	14.151	11.293

Se observa que los valores calculados de RE_i se corresponden con sus respectivos ECM_i. Por ejemplo la variable 3 tiene un RE_i próximo a 1 y el menor ECMi, esto implica que el procedimiento de RDR logra una mayor corrección de las mediciones de esta variable. En el otro extremo se encuentra la variable 1 con el menor valor de redundancia y mayor ECM_i. Las mismas tendencias de RE_i y ECM_i se observan si se ejecuta el procedimiento de RDC. Con este ejemplo se demuestra que la expresión obtenida por Maronna puede utilizarse en sistemas de ecuaciones linealizados.

4.6 Análisis Exhaustivo del Desempeño de los Test Estadísticos

A continuación se presentan los resultados del análisis de desempeño de los test estadísticos propuestos en esta tesis.

4.6.1 Medidas de desempeño

Narasimhan y Mah (1987) presentaron dos medidas para evaluar el desempeño del Test de Razón de Máxima Verosimilitud en presencia de múltiples ESE. Éstas son el

Desempeño Global, conocido por sus siglas en inglés como OP (Ec. 3.43), y el Número Promedio de Errores Tipo 1, conocido por sus siglas en inglés como AVTI (Ec. 3.44). El primero cuantifica la capacidad de detectar errores y el segundo está relacionado a las falsas alarmas del test. Dichas métricas son adecuadas si las hipótesis del test estadístico son iguales a las vistas para el TM. En tal caso, H_0 se refiere a la ausencia de error sistemático en el vector de medición, por esto el AVTI está relacionado con la cantidad de simulaciones de dicho vector.

Por el contrario, la H_0 del TRM se refiere a la ausencia de error en cada medición del vector de observaciones actual, por lo que las medidas de desempeño anteriores no son adecuadas para evaluar el comportamiento del nuevo test. En consecuencia, se define el Porcentaje de Falsas Alarmas de ESE ($\%FA_{ESE}$) como medida del Error Tipo 1 (ET1) del TRM. El $\%FA_{ESE}$ se calcula con la siguiente expresión.

$$\%\mathrm{FA}_{\mathrm{ESE}} = \frac{(ESE)_{Detectados} - (ESE)_{Simulados\ y\ Detectados}}{I\ Ns}\, x\, 100 \tag{4.47}$$

Por su parte, el OP se particulariza en este capítulo para los ESE y se lo reformula como:

$$\%\mathrm{DT}_{\mathrm{ESE}} = \frac{(ESE)_{Identificados}}{(ESE)_{Simulados}}\, x\, 100 \tag{4.48}$$

Algunas generalidades para las subsiguientes pruebas son:

- Se realizan 10.000 simulaciones para obtener medidas de desempeño reproducibles;
- Se emplea el Modelo 3 (Sección 3.6) para la generación de mediciones atípicas;
- El estadístico crítico del TRM se calcula considerando $\alpha = 0,025$ y el *df* depende del tamaño de la ventana;

- La probabilidad global (p_G) es la probabilidad de presencia de ESE en el conjunto total de mediciones simuladas;
- La probabilidad individual (p_I) es la probabilidad de presencia de ESE en una variable medida en particular;
- El desvío estándar de las mediciones es 2,5% de su valor verdadero para todos los casos de estudio representados por sistemas lineales.

4.6.2 Prueba 1: Comparación del Test TMV y TRM

La evaluación del comportamiento de los TRM y TMV se efectúa con el objetivo de mostrar las mejoras conseguidas con la utilización de conceptos de la Estadística Robusta. El TRM calcula el $\mathbf{a^R}$ con el $\hat{\mathbf{x}}$ de la metodología MSi, en cambio el TMV utiliza la función CM inicializada con la media de la ventana de datos.

Como caso de estudio se emplea el ejemplo propuesto por Rosemberg (1987), que se presenta en la Fig. 4.1. El sistema lineal comprende 4 ecuaciones de balance y 7 variables medidas. Se realizan 10.000 simulaciones para longitudes de ventana N=10, 20, 30 y 40. Se emplea el mismo conjunto de datos para ejecutar ambas metodologías. La p_G de ESE se fija en 10% y la magnitud de ESE se varía en el rango [0-20].

El τ_c^R del TRM corresponde a un nivel de significancia α=0,025 y varía con N, pues $\tau_c^R = t_{N-1,1-\alpha/2}$. Para realizar una comparación acertada se variará el parámetro α del TMV de manera que ambos test tengan la misma cantidad de falsas alarmas cuando sólo hay errores aleatorios. Las medidas de desempeño que se utilizan son: ECM, $\%DT_{ESE}$ y $\%FA_{ESE}$ y los resultados obtenidos se presentan en la Fig. 4.2.

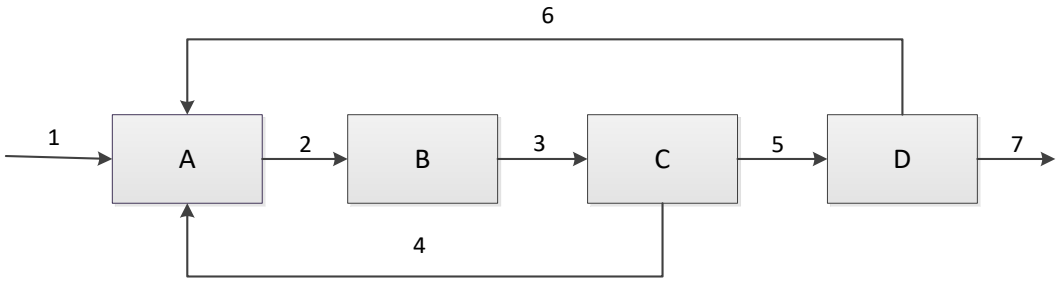


Figura 4.1 Sistema Lineal de Rosenberg (1987).

El análisis comparativo de las medidas desempeño muestra que:

- El ECM disminuye a medida que *N* aumenta. Esto indica que ambas metodologías de reconciliación se ven favorecidas por la RT, en especial cuando *N*>10.
- El ECM de la metodología robusta es acotado e inferior al de la metodología que empleada la función CM, esta diferencia se acrecienta a medida que los ESE simulados son de mayor magnitud. Básicamente, la Estadística Robusta pondera a las mediciones otorgándole menor peso a aquellas cuyos ajustes se encuentran ubicados en las colas de la distribución normal. Por esto cuando *K*> 6, el ECM se mantiene prácticamente constante mientras que el ECM correspondiente al TVM aumenta de forma monótona.
- El TMV tiene mayor $\%DT_{ESE}$ que el TRM en todas las condiciones analizadas. El TMV consigue $\%DT_{ESE}$ elevados e idénticos para todas las longitudes de ventana, por el contrario el TRM consigue mejorar su desempeño a mayores *N*. Esto se debe a que el TRM estima una varianza muestral del $\mathbf{a}^{\mathbf{R}}$ la cual es más precisa cuando el conjunto de datos considerado en el cálculo es mayor.
- El TMV consigue detectar el 99% de los ESE de magnitud igual o mayor que $6\sigma_{y,}$ mientras que el TRM detecta el 96 % de esos errores con *N* iguales a 30 o 40. Los ESE de mayor magnitud logran ser detectados por ambos métodos con igual eficacia.

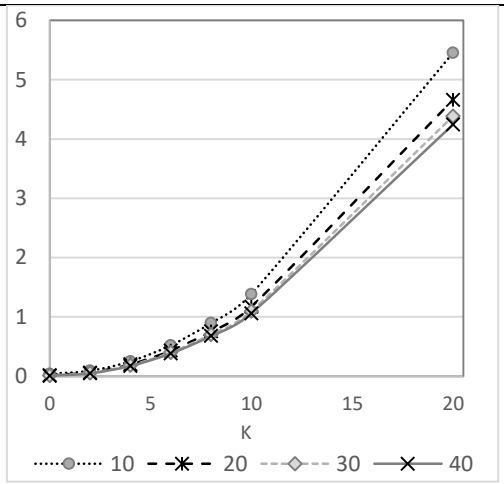
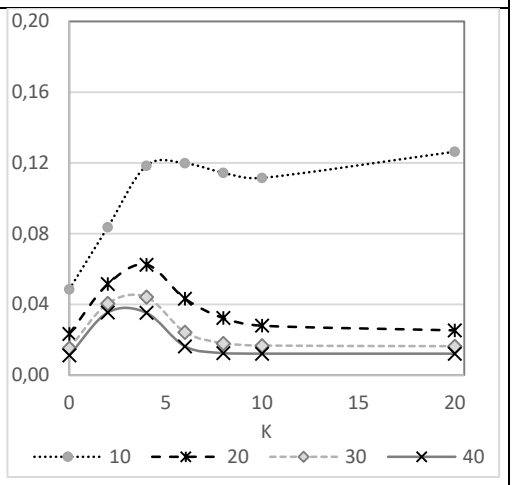
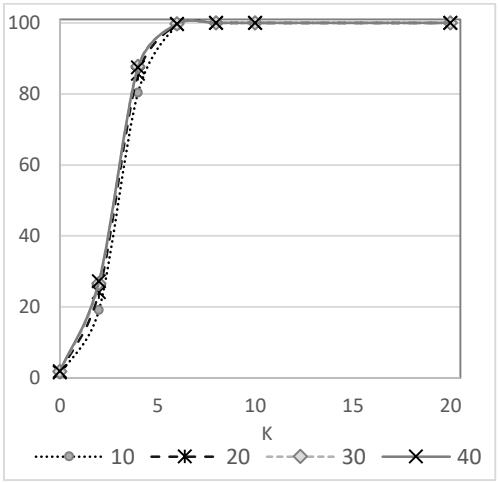
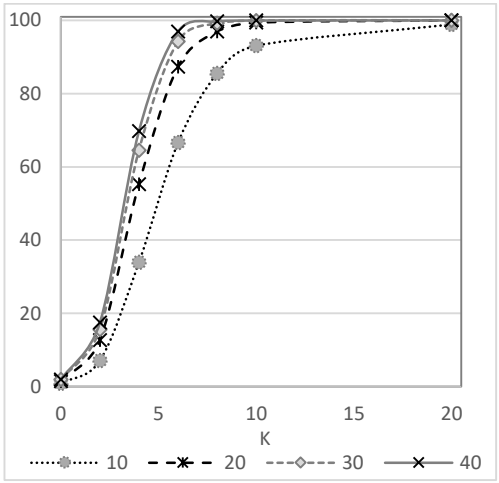
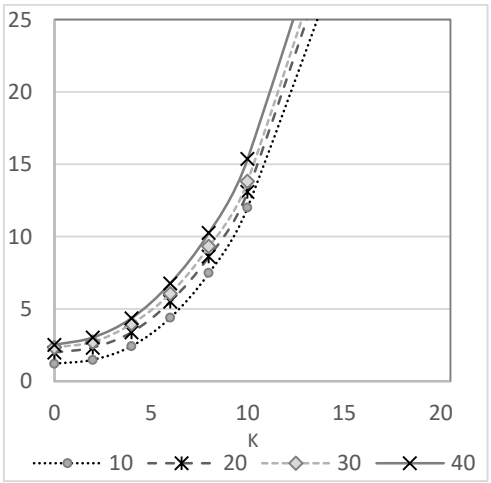
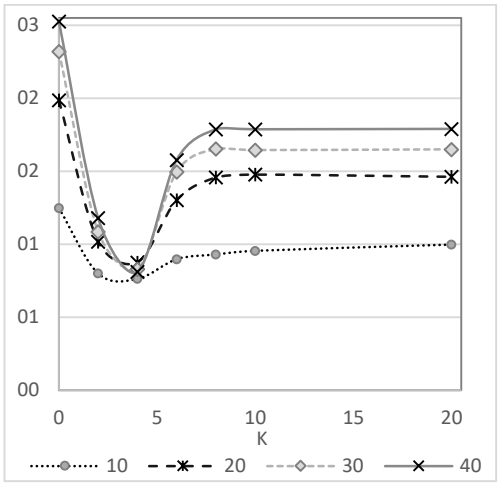
ÍNDICE	TMV	TRM
ECM		
	Fig 4.2.a	Fig 4.2.d
%DT$_{ESE}$		
	Fig 4.2.b	Fig 4.2.e
%FA$_{ESE}$		
	Fig.4.2.c	Fig 4.2.f

Figura 4.2. Curvas de desempeño del TVM y el TRM para *N* [10-40]

- El TMV tiene curvas de $\%FA_{ESE}$ crecientes con la magnitud del ESE. Esto se debe a que si la magnitud del error es grande hay mayor efecto de dispersión del error entre las estimaciones de las variables. También se observa un incremento del ET1 con el tamaño de la ventana, aunque este es menos marcado.
- Las curvas de $\%FA_{ESE}$ son inicialmente convexas para el TRM y luego prácticamente constantes. Esto se debe a que el test tiene capacidad limitada para detectar errores de magnitud inferiores a 4, mientras que para $K>6$ el $\%DT_{ESE}$ alcanza el máximo y las falsas alarmas permanecen prácticamente constantes.
- Los $\%FA_{ESE}$ del TRM aumentan con N. Esto se debe a que el valor del t_c disminuye a medida que aumentan los df, con lo cual una mayor cantidad de estadísticos pueden superar dicho valor. No obstante el $\%FA_{ESE}$ es siempre inferior al α propuesto.

Este análisis permite concluir que los resultados de la RDC basada en la función CM se ven afectados por la presencia de ESE, esto repercute sobre el $\mathbf{a}$ generando identificaciones incorrectas de errores sistemáticos. Por el contrario, el TRM formula el $\mathbf{a}^R$ utilizando el valor estimado dado por la metodología MSi. Ésto origina que el $\mathbf{a}^R$ sea insensible a la presencia de ESE, con lo cual los efectos de dispersión del error se minimizan. El TRM muestra una mejor relación entre el $\%DT_{ESE}$ y el $\%FA_{ESE}$ lo cual lo hace útil para la Detección de ESE (DES).

4.6.3 Prueba 2: Influencia de la redundancia temporal en el TRM

Como se ha mencionado, la metodología propuesta por Maronna y Arcas (2009) permite obtener las RE_i. La información provista por esta medida es útil para analizar el comportamiento de una técnica de DES, pues su desempeño disminuye para las variables con baja RE_i. Sin embargo, la RT puede contribuir a mejorar el desempeño del test. Por esto se proponen las siguientes pruebas para el mismo caso de estudio:

Prueba 2.1: Se analiza el comportamiento del TRM cuando sólo existe RE. Para esto se omite el primer paso de la metodología MSi. Cabe aclarar que para el cálculo de la matriz de covarianza se utiliza una ventana de ajustes correspondiente a N=40.

Prueba 2.2: Se analiza el comportamiento del TRM cuando el procedimiento de RDR hace uso de la RT provista por una ventana de mediciones de N=40.

Se generan mediciones atípicas con una p_I = 0,1; esto se efectúa para analizar el comportamiento de cada variable en función de su respectiva RE_i. La Prueba 2.1 mostró que no existen diferencias significativas en el comportamiento del TRM para ESE con K >8, por lo que se realizan simulaciones en el rango K = [0 - 8]. Se presentan las RE_i calculadas siguiendo el procedimiento de Maronna y Arcas (2009) en la Tabla 4.2, y las medidas de desempeño obtenidas para las dos pruebas en la Fig. 4.3. Dado que las variables 2 y 3 tienen igual RE_i, y lo mismo sucede con las variables 1 y 7, sólo se grafican las curvas de desempeño de las variables 1, 2, 4, 5 y 6.

Tabla 4.2 Valores de RE_i de las variables

Variable	RE_i	Variable	RE_i	Variable	RE_i	Variable	RE_i
1	0,57	**3**	0,84	**5**	0,77	**7**	0,57
2	0,84	**4**	0,15	**6**	0,26		

1) Prueba 2.1

- Las curvas de $\%DT_{ESE}$ y las de ECM_i presentan los resultados esperados según la teoría de DES. A las variables con RE_i próximas a 1 les corresponden curvas con menor ECM_i, lo contrario sucede con aquellas variables que tienen RE_i tendiendo a 0. Esto se debe a que las variables del primer grupo pueden ser más corregidas por la RD y por lo tanto ser más próximas a sus valores verdaderos.

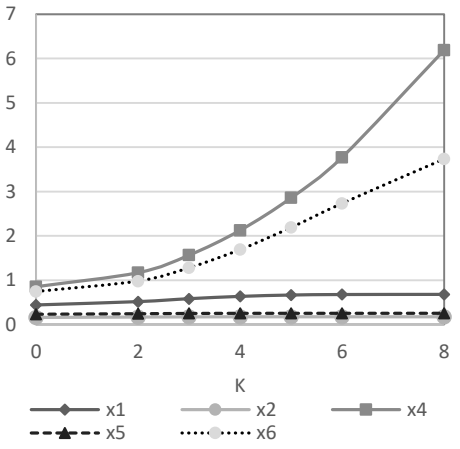
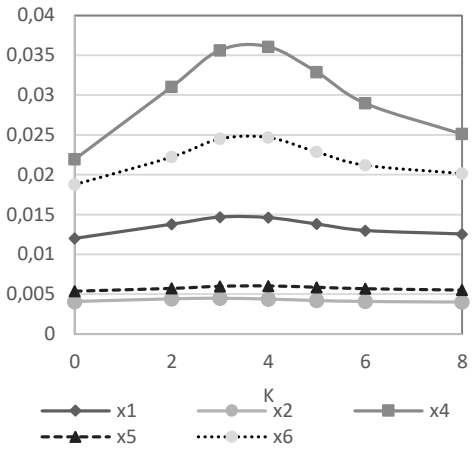
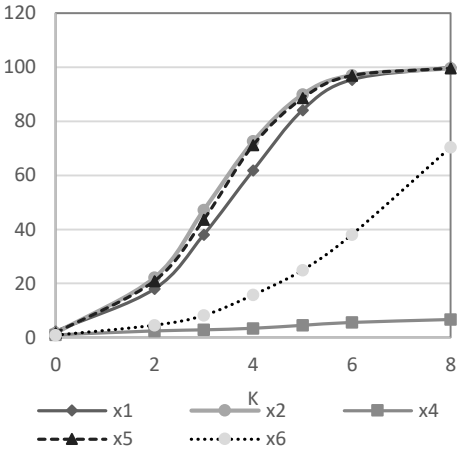
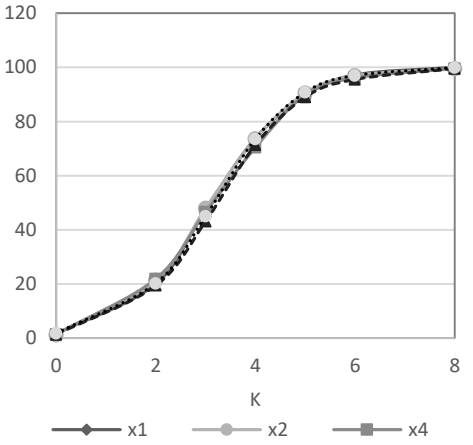
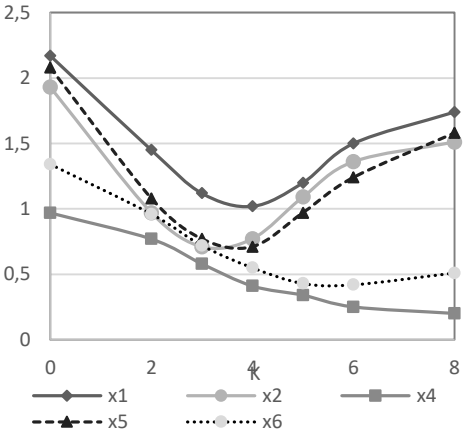
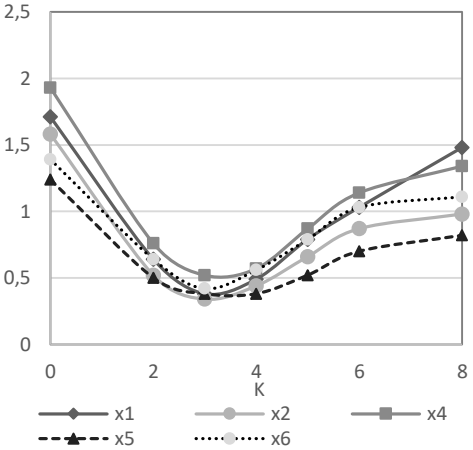
ÍNDICE	Prueba 2.1	Prueba 2.2
ECMi	 Fig 4.3.a	 Fig 4.3.d
%DTESE	 Fig 4.3.b	 Fig 4.3.e
%FAESE	 Fig.4.3.c	 Fig 4.3.f

Figura 4.3 Curvas de Desempeño de las Pruebas 2.1 y 2.2

- Las Curvas de $\%FA_{ESE}$ muestran que las variables con baja redundancia tienen menores porcentajes de FA. Esto se debe a que la capacidad de detección del TRM en estas condiciones también es escasa. Cabe aclarar que a diferencia del TMV de la Prueba 1, que emplea la función CM; todas las curvas de $\%FA_{ESE}$ tienen valores inferiores a 2.5 %.

2) Prueba 2.2

- Las curvas de ECM_i reflejan la influencia de la RE_i. A pesar de notarse la influencia de la RT, se observa que la curva que corresponde a los valores máximos de ECM_i se obtiene para la variable con menor RE_i; lo contrario sucede en la variable con mayor RE_i.
- Todas las variables tienen igual capacidad para detectar ESE. La RT mejora los resultados de la RDR y esto favorece al TRM. No se observan diferencias apreciables entre las curvas de $\%FA_{ESE}$.

El uso de la RT, utilizada en el primer paso de la metodología MSi, permite la detección de mediciones en todas las variables con ESE. El uso de una metodología como MSi provee de estimaciones exactas que otorgan al TRM elevada capacidad de detección, independientemente de la baja RE dada por el modelo del proceso. A continuación se analiza que sucede cuando algunas variables tienen RE_i nula.

4.6.4 Prueba 3: Desempeño del TRM en Variables No Redundantes

Las pruebas también se realizan sobre el diagrama de flujo propuesto por Rosenberg, pero se asume que sólo se dispone de 4 mediciones (Fig. 4.4). La Tabla 4.3 presenta las RE_i. Se procede de manera similar a la Prueba 2, es decir, se analiza cada

variable en particular y se trazan las correspondientes curvas de desempeño. Los ESE se generan aleatoriamente con una $p_I = 0.1$ y se utilizan ventanas de $N = 40$.

En la Fig. 4.5, se observa que las curvas de $\%DT_{ESE}$ y $\%FA_{ESE}$ para la variable 1 (redundante) y para la variable 4 (no redundante) presentan resultados similares. Se concluye que el desempeño del test resulta elevado aun en variables no redundantes. Estas pruebas se realizaron con una p_I de ocurrencia de error del 10%, la cual corresponde a una p_G del 2,5%. A continuación se analiza el deterioro del test a mayores p_G.

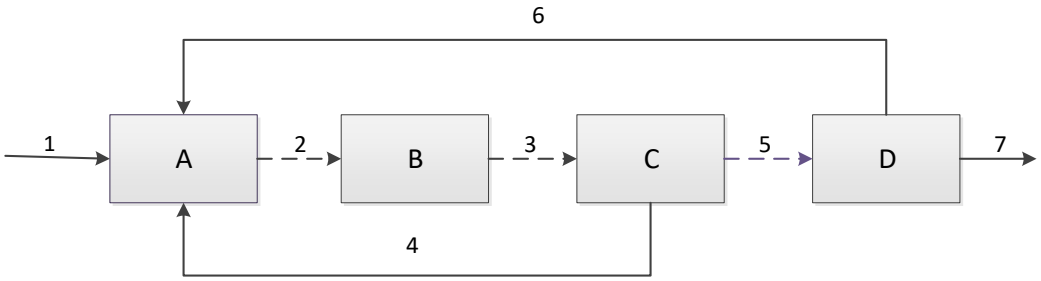


Figura 4.4. Proceso de Rosenberg con variables no medidas

Tabla 4.3. Redundancia de las variables medidas

Variable	1	4	6	7
RE_i	0,5	0	0	0,5

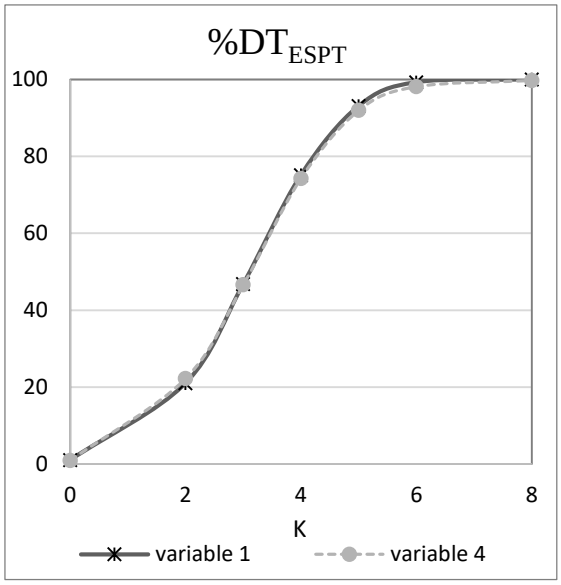


Figura 4.5.a Curva de $\%DT_{ESPT}$

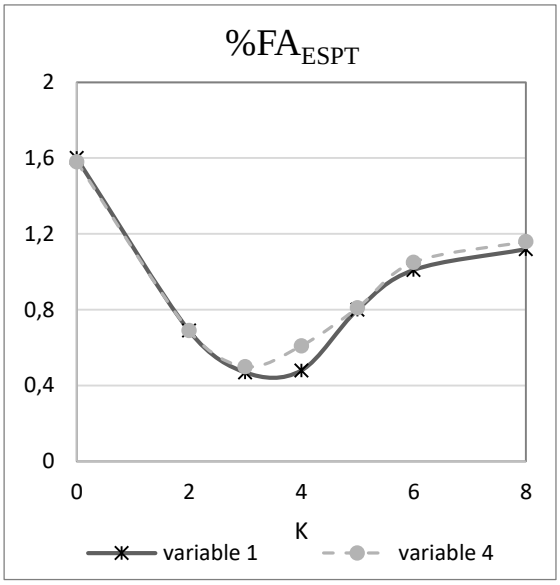


Figura 4.5.b Curva de $\%FA_{ESPT}$

4.6.5. Prueba 4: Desempeño del TRM para distintas p_G

Se evalúa el desempeño del test para distintas p_G de ESE. Para este análisis se utiliza el diagrama de flujo propuesto por Rosenberg en la condición menos favorable, o sea con 3 variables no medidas (Fig 4.4). Se generan ESE con p_G en el rango [0,1-0,3] siendo la longitud de la ventana *N*=40. En la Tabla 4.4 se presenta la cantidad de ESE simulados cuando se generan 40000 mediciones.

Tabla 4.4. Cantidad total de ESE simulados

	$P_G = 0,1$	$P_G = 0,15$	$P_G = 0,2$	$P_G = 0,25$	$P_G = 0,3$
Número Errores Generados	4046	6023	7931	10034	12023

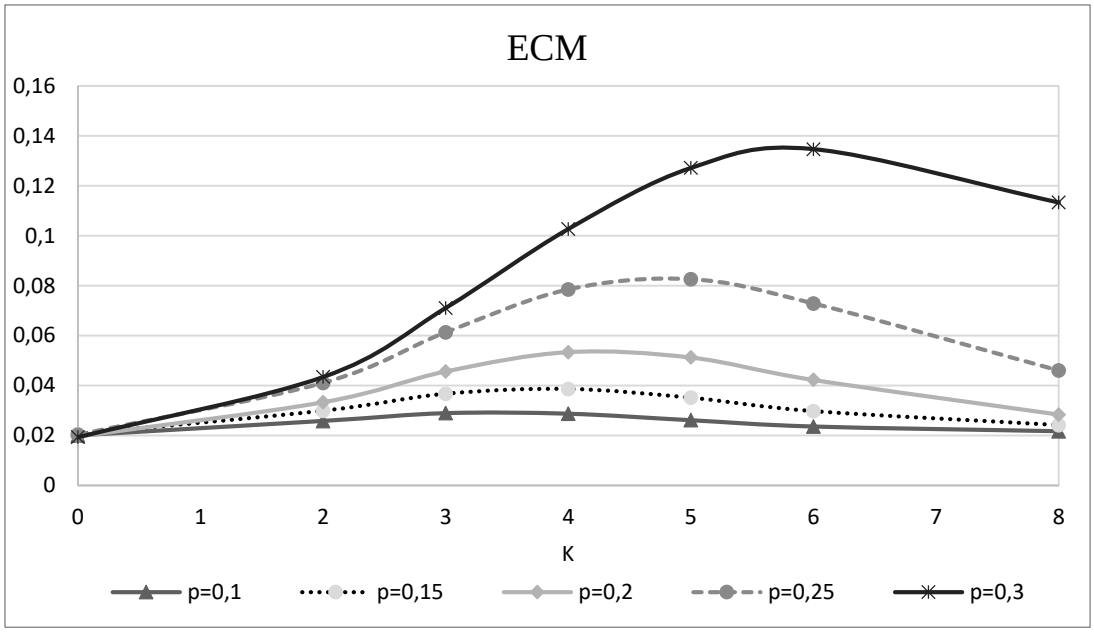


Figura 4.6a Curvas de ECM para el proceso de la Figura 4.4 y distintos p_G

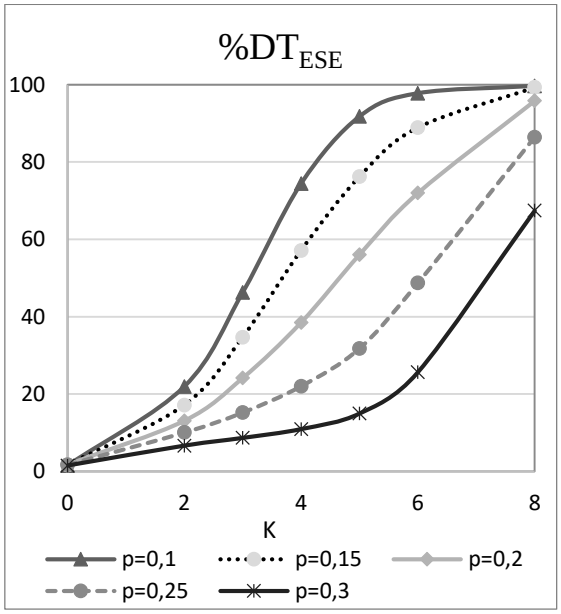


Figura 4.6.b Curvas de detección de ESE a distintas p_G

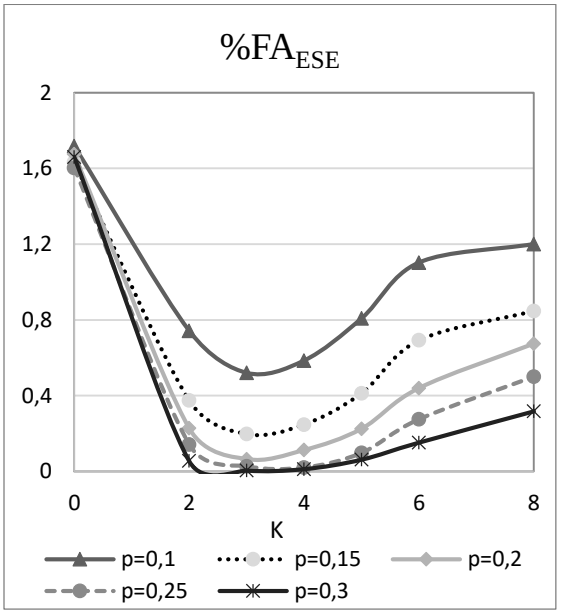


Figura 4.6.c Curvas de falsas alarmas de ESE a distintos p_G

En las Figuras 4.6 a,b y c se observa que:

- El ECM aumenta con la p_G de ESE. Esto se debe a que la RDR es insensible sólo a una cantidad limitada de errores sistemáticos. La primera etapa de MSi calcula la mediana robusta. El incremento de la p_G produce un aumento en la probabilidad de tener mediciones atípicas consecutivas, lo cual perjudica a la mediana robusta y al procedimiento de RDR.
- El $\%DT_{ESE}$ disminuye con el aumento de la p_G. Las estimaciones resultantes del procedimiento de RDR son menos exactas cuando se tiene una mayor cantidad de mediciones atípicas en una ventana de datos. Esto afecta al $\mathbf{a}^{\mathbf{R}}$ deteriorando el desempeño del test.
- El $\%FA_{ESE}$ disminuye con la p_G. Al aumentar la cantidad de errores simulados aumentan las coincidencias entre mediciones atípicas y mediciones que se encuentran en las colas de la distribución normal.

Los parámetros de desempeño muestran que el comportamiento de la metodología MSi seguida del TRM da bueno resultados para p_G de hasta 0,2. Luego el ECM aumenta al doble, con lo cual se tienen estimaciones menos precisas y disminuye el desempeño del TRM.

4.6.6 Prueba 5: Sistemas no lineales con redundancia baja y nula

Se analiza el desempeño del test en sistemas no lineales que comprenden variables medidas de distinto grado de redundancia y variables no medidas. Al igual que en los sistemas lineales el procedimiento de RDR se realiza con la metodología MSi. La segunda etapa de este procedimiento se ejecuta con el paquete de optimización de Matlab (2014).

Se utilizan como casos de estudio las ecuaciones de Pai y Fisher (1988), P&F, y una Red de Intercambiadores de Calor, HEN (Swartz, 1989). Esta última se presenta en

la Fig. 4.7; involucra 15 corrientes que intercambian calor y comprende 30 variables (15 caudales másicos y 15 temperaturas) de las cuales 16 son medidas y 14 no medidas. Se formulan 17 balances de masa y energía alrededor de los intercambiadores de calor, mezcladores y divisores. Se asume que el desvío estándar de las temperaturas es 0,75K y los desvíos estándares de los caudales son el 2% de sus valores verdaderos.

En la Tabla 4.5 se reportan los valores de RE_i para las variables del caso de estudio HEN obtenidos a partir del desarrollo presentado en la Sección 4.5.2, mientras que en la Fig. 4.8 se presentan las medidas de desempeño cuando la p_I de ESE es del 10% y N = 40.

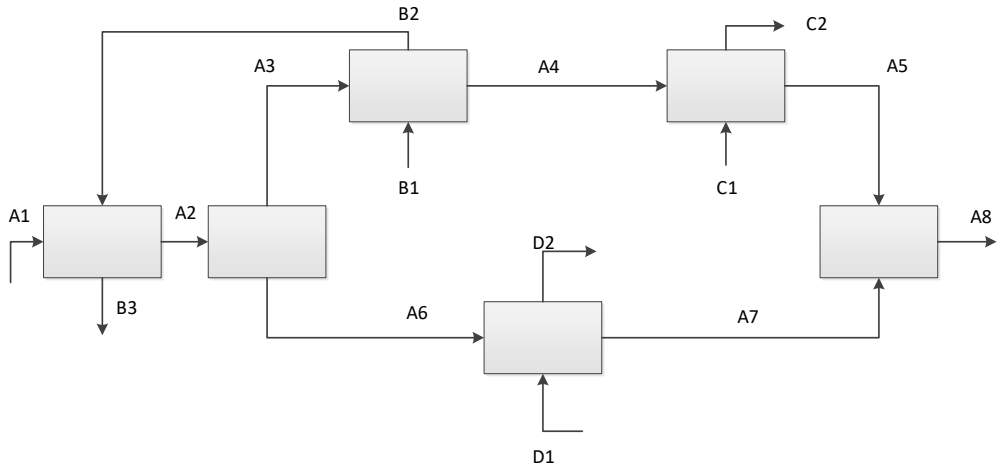


Figura 4.7. Red de Intercambio de Calor (HEN)

Tabla 4.5 Redundancia de las variables medidas (HEN)

VARIABLE	Nº	RE_i	Nº	VARIABLE	RE_i
FA1	1	0,75	TB1	19	0
FA3	3	0,12	TA4	21	0
FB1	4	0	TC1	23	0
FC1	8	0	TA5	25	0,12
FA6	12	0,50	TA8	26	0,67
FD2	14	0,46	TD1	28	0,06
TA1	16	0	TD2	29	0,05

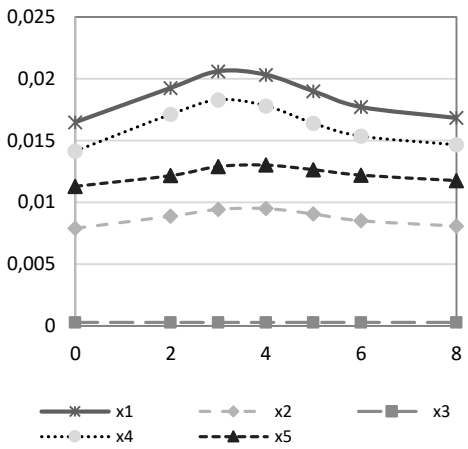
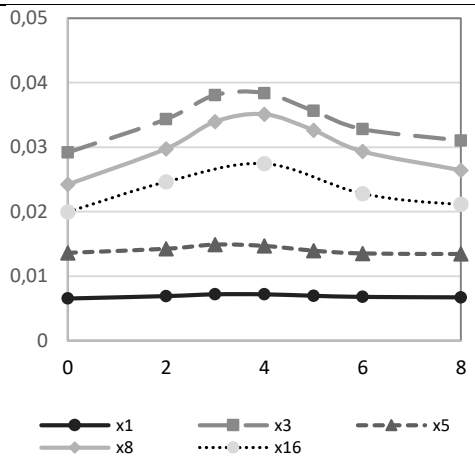
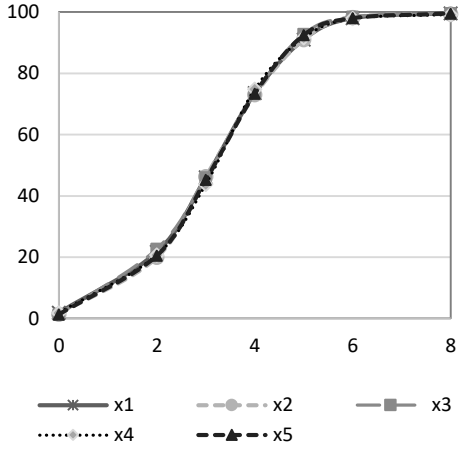
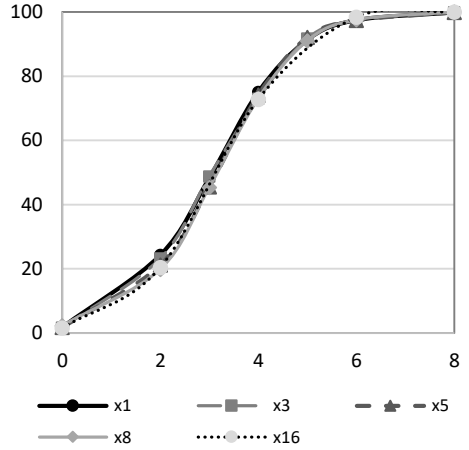
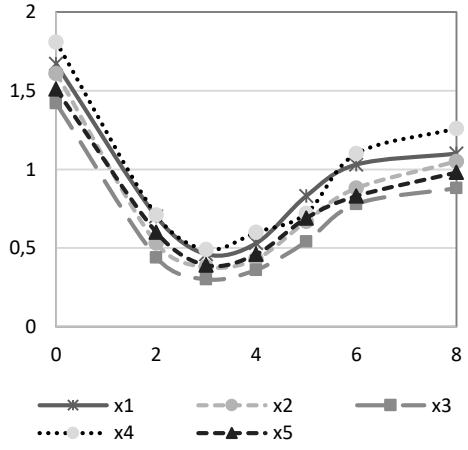
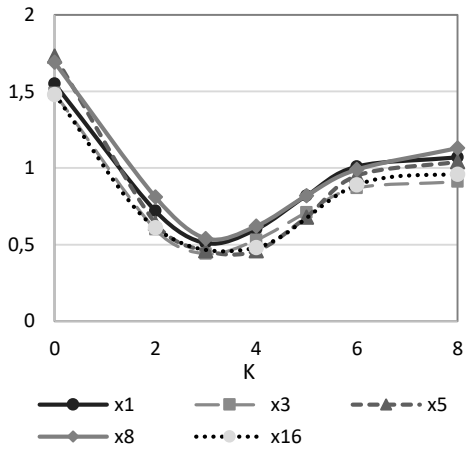
ÍNDICE	P&F	HEN
ECM_i	 Fig 4.8.a	 Fig 4.8.d
$\%DT_{ESE}$	 Fig 4.8.b	 Fig 4.8.e
$\%FA_{ESE}$	 Fig.4.8.c	 Fig 4.8.f

Figura 4.8 Curvas de desempeño para sistemas no lineales

El análisis de las gráficas para los dos sistemas muestra que:

- Las curvas del ECM se condicen con las RE_i calculadas. Al igual que en los sistemas lineales, las curvas con menor RE_i tienen mayores valores de ECM_i.
- Los $\%DT_{ESE}$ y $\%FA_{ESE}$ son similares para todas las variables medidas. La RT mejora la capacidad de detección de las variables con baja RE.

Por lo que se comprueba que el desempeño del TRM en sistemas no lineales es similar al obtenido para sistemas lineales.

4.6.7 Desempeño del TRM en sistemas con problemas estructurales

Rosenberg y co. (1987) mencionan que son varios los factores que pueden interferir en el desempeño de una metodología de DES y los clasifica en cuatro categorías. La primera está relacionada a la magnitud del error; la segunda a la redundancia de las variables, la tercera a los diferentes órdenes de magnitud que pueden tener las variables de un sistema y la última a las restricciones y estructura de los modelos. Las primeras tres categorías han sido analizadas en las pruebas anteriores, por lo que resta analizar el comportamiento del test para la cuarta categoría.

Iordache y co. (1985) mostraron que los estadísticos del TM para dos mediciones diferentes son iguales si sus correspondientes columnas en la matriz que representa el modelo lineal son proporcionales. Esto imposibilita distinguir la variable con ESE. Un caso particular de esto sucede cuando las corrientes son paralelas, o sea que unen los mismos dos nodos del proceso. Narasimhan y Mah (1987) también detectaron ésta situación utilizando el TRMV. Años más tarde Sánchez (1996) estableció las condiciones bajo las cuales se presentaban los problemas estructurales, y Bagajewicz y Jiang (1999) propusieron el concepto de conjuntos equivalentes de errores para referirse a la imposibilidad de identificar un error cuando éste se encuentra en determinadas corrientes.

Romagnoli y Sánchez (2000) mostraron el desempeño de tres metodologías de DES cuando los errores se encuentran en variables equivalentes empleando el ejemplo de Rosenberg (Fig 4.1). Bagajewicz (2010) reporta que este problema aún no ha logrado ser resuelto.

A continuación se analiza el desempeño del TRM en casos de estudio que presentaron dichos problemas estructurales. Se realizan 10.000 simulaciones y se colocan ESE en las variables reportadas como no identificables.

4.6.7.1 Prueba 6: Columnas proporcionales

El diagrama de flujo presentado en la Fig 4.9 se denota CP1987 y fue propuesto por Rosenberg y co. (1987). El modelo del proceso comprende 4 ecuaciones de balance y 7 variables medidas, y la matriz asociada al sistema lineal de ecuaciones se incluye en la Fig 4.10. Se observa que la matriz presenta dos columnas idénticas que corresponden a las variables con problema de identificación.

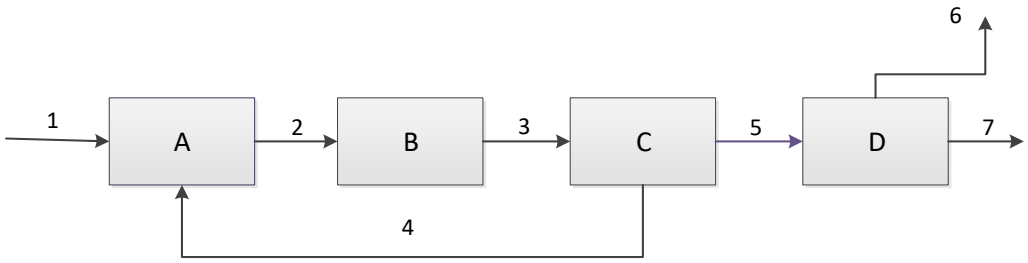


Figura 4.9. Proceso del ejemplo CP1987

$$\mathbf{A}_1 = \begin{bmatrix} 1 & -1 & & 1 & & & \\ & 1 & -1 & & & & \\ & & 1 & -1 & -1 & & \\ & & & & 1 & -1 & -1 \end{bmatrix}$$

Figura 4.10 Matriz del modelo lineal del ejemplo CP 1987

Se realizan 2 pruebas para determinar la capacidad del TRM para detectar ESE en dichas variables medidas.

Prueba 6.1: Se colocan ESE en la variable 6 y se evalúa el $\%DT_{ESE}$, $\%FA_{ESE}$ y las falsas alarmas en la variable 7 ($\%\ FA_7$).

Prueba 6.2: Se colocan ESE en la variable 7 y se evalúa el $\%DT_{ESE}$, $\%FA_{ESE}$ y las falsas alarmas en la variable 6 ($\%\ FA_6$).

Los resultados alcanzados al realizar la Prueba 6.1 y la Prueba 6.2 se muestran en las Tablas 4.6 y 4.7, respectivamente

Tabla 4.6 Resultados de la Prueba 6.1

K	$\%DT_{ESE}$	$\%FA_{ESE}$	$\%FA_7$
0	1,95	1,60	1,60
2	20,64	1,49	1,62
3	45,28	1,47	1,61
4	74,49	1,48	1,62
5	91,92	1,50	1,60
6	97,57	1,53	1,57
8	100,00	1,54	1,56

Tabla 4.7 Resultados de la Prueba 6.2

K	$\%DT_{ESE}$	$\%FA_{ESE}$	%FA6
0	1,56	1,61	1,64
2	20,06	1,49	1,64
3	45,76	1,47	1,64
4	74,20	1,49	1,64
5	92,31	1,51	1,65
6	98,54	1,52	1,65
8	99,90	1,53	1,65

Del análisis de las tablas previas surge que es posible identificar correctamente la variable con ESE sin que se produzca un aumento de $\%FA_{ESE}$. Los $\%FA_6$ y $\%FA_7$ no presentan cambios significativos.

4.6.7.2 Prueba 7: Corrientes paralelas

Romagnoli y Stephanopoulos (1981) presentaron el diagrama de flujo de un proceso con corrientes paralelas (Fig 4.11), que está formado por tres equipos que interconectan 6 variables medidas. La matriz asociada al sistema lineal de ecuaciones se incluye en la Fig 4.12. Este ejemplo se denota como R&S. Se colocan ESE en las mediciones correspondientes a las corrientes paralelas y se miden los parámetros de desempeño para distinta magnitudes de ESE. Las pruebas realizadas son las siguientes:

Prueba 7.1: Se colocan ESE en la variable 2 y se evalúa el $\%DT_{ESE}$, $\%FA_{ESE}$ y las falsas alarmas en la variable 3 ($\%\ FA_3$).

Prueba 7.2: Se colocan ESE en la variable 3 y se evalúa el $\%DT_{ESE}$, $\%FA_{ESE}$ y las falsas alarmas en la variable 6 ($\%\ FA_2$).

Los resultados alcanzados al realizar la Prueba 7.1 y la Prueba 7.2 se muestran en las Tablas 4.8 y 4.9, respectivamente.

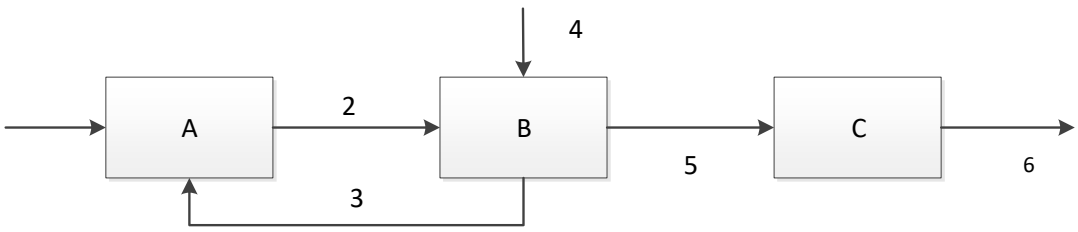


Figura 4.11. Sistema con corrientes paralelas (R&S)

$$A = \begin{bmatrix} 1 & -1 & 1 & & & \\ & 1 & -1 & 1 & -1 & \\ & & & & 1 & -1 \end{bmatrix}$$

Figura 4.12. Matriz del modelo lineal del ejemplo R&S

Tabla 4.8: Resultados de la Prueba 7.1

K	$\%DT_{ESE}$	$\%FA_{ESE}$	$\%FA_3$
0	1,46	1,37	1,63
2	18,45	1,28	1,75
3	43,30	1,24	1,75
4	72,82	1,24	1,74
5	90,97	1,27	1,70
6	97,77	1,29	1,70
8	99,90	1,31	1,69

Tabla 4.9: Resultados de la Prueba 7.2

K	$\%DT_{ESE}$	$\%FA_{ESE}$	$\%FA_2$
0	1,36	1,37	1,41
2	22,14	1,28	1,41
3	42,72	1,25	1,38
4	72,52	1,24	1,36
5	90,39	1,29	1,35
6	97,28	1,31	1,36
8	99,90	1,33	1,37

Las conclusiones son similares a las reportadas para la Prueba 6.

4.6.7.3 Prueba 8: Corrientes equivalentes

Romagnoli y Sánchez (2000) presentaron la pérdida de desempeño de las metodologías SEGE, TRMV y TMIM cuando los ESE se colocan en 2 variables correspondientes al

conjunto de corrientes equivalentes del caso de estudio propuesto por Rosemberg y col. (1987). En esas condiciones los autores observaron un aumento de falsas alarmas; éstas se daban en la variable equivalente sin ESE.

A continuación se evalúa el desempeño del TRM para el mismo escenario, con una p_I = 0,1 y ESE de magnitudes K=7 y 4 en cada par de variables. Las medidas de desempeño ($\%DT_{ESE}$ y $\%FA_{ESE}$) se presentan en la Tabla 4.10 y se agrega un tercer parámetro que evalúa el porcentaje de falsas alarmas en la corriente equivalente sin ESE, $\%FA_{EQ}$.

Tabla 4.10. Resultados de la Prueba 8

CAUDAL CON ESE	$\%DT_{ESE}$	$\%FA_{ESE}$	$\%FA_{EQ}$
1—2	87,08	1,44	-
1—3	86,49	1,42	-
1—4	86,73	1,44	-
1—5	87,08	1,41	-
1—6*	86,19	1,41	1,64(7)
1—7*	86,09	1,46	1,76(6)
2—3*	86,49	1,43	1,72(4)
2—4*	86,68	1,45	1,78(3)
2—5	86,68	1,45	-
2—6	86,14	1,43	-
2—7	86,19	1,46	-
3—4*	86,73	1,44	1,53(2)
3—5	87,18	1,41	-
3—6	86,24	1,43	-
3—7	86,34	1,45	-
4—5*	87,23	1,42	1,74(6)
4—6*	86,34	1,43	1,66(5)
4—7	86,34	1,46	-
5—6	86,19	1,44	1,71(4)
5—7	86,14	1,45	-
6—7*	86,39	1,44	1,59(1)

*Variables equivalentes, ( ) variable equivalente sin ESE

En la Tabla 4.10 se observa que:

- Las medidas de desempeño de todos los pares analizados no muestran diferencias significativas (ds_%DT=0,36 y ds_%FA=0,016).
- El $\%FA_{EQ}$ es superior al %FA. Según lo reportado por Romagnoli y Sánchez (2000) la variable equivalente sin ESE da FA de error cuando su par equivalente tiene ESE. Con lo cual, para una probabilidad de 0.1 en un par de errores las FA_{EQ} deberían dar próximas al 10%. Por esto se considera que el aumento de $\%FA_{EQ}$ es insignificante en comparación al dado por otras metodologías.

4.7 Conclusiones

En este capítulo se presentan metodologías para detección e identificación simultánea de ESE, que hacen uso de la RT de las mediciones contenidas en una ventana de datos. Estos test se denominan TMV y el TRM, y a diferencia de metodologías previas, permiten detectar e identificar ESE tanto en mediciones redundantes como en las no redundantes.

La comparación de las medidas de desempeño del TMV y TRM exhibe la pérdida de exactitud en las estimaciones de la RDC en relación con la provista por la RDR. Aunque el TMV logra mayor detección de ESE, lo hace a expensas de un gran número de falsas alarmas, en cambio, el TRM no sufre incrementos en el %FA por la presencia de mediciones atípicas. Además, si bien el TRM consigue detectar un menor porcentaje de ESE en errores de baja magnitud, esto no interfiere en los resultados de la RDR. Por otra parte, una mayor longitud de la ventana de datos mejora el desempeño del TRM.

Los resultados muestran la relación existente entre el ECM y la RE, confirmando que la detección de ESE es menor en las variables con poca redundancia. Sin embargo el

uso de un conjunto de datos presentes en una ventana provee de RT, con la cual se consiguen $\%DT_{ESE}$ y $\%FA_{ESE}$ similares para todas las variables independientemente de la escasa o nula RE de las mismas. Esto es un notable avance en las técnicas de DES pues independiza la capacidad de detección de la RE.

En relación con la aplicación de las metodologías a sistemas no lineales, el desarrollo de Maronna y Arcas (2000) permite obtener estimaciones de la RE de las variables involucradas en estos sistemas y además, el TRM tiene medidas de desempeño idénticas a las correspondientes a sistemas lineales.

Además, el TRM permite identificar las variables con ESE en sistemas complejos, como los que presentan corrientes paralelas o conjuntos equivalentes de mediciones, sin producir incremento en las falsas alarmas; con lo cual se aborda un problema cuya solución estaba pendiente hasta el momento.

Dado que el TRM permite identificar mediciones atípicas en todas las variables medidas independientemente de la RE de éstas para procesos representados tanto por ecuaciones algebraicas lineales como no lineales con bajas falsas alarmas, resulta útil para desarrollar estrategias capaces de detectar errores que persisten en el tiempo. Este tema se aborda en el Capítulo 5 de esta tesis.

4.8 Notación

a	Vector de ajuste de las mediciones
$\mathbf{A_1}$	Matriz representativa de las ecuaciones lineales de reconciliación
$\mathbf{A^R}$	Matriz de ajustes
Ajus	Ajustabilidad
Det	Determinabilidad
I	Número de variables medidas
K	Magnitud del ESE
M	Número de ecuaciones del modelo
N	Número de réplicas de la variable medida
Ns	Número de Simulaciones
Q	Matriz de covarianza del ajuste
Q_d	Matriz de covarianza del ajuste estandarizado
Y	Vector de mediciones
X	Vector de variables medidas
$\hat{\mathbf{x}}$	Vector reconciliado de las variables medidas
Y_{ob}	Matriz de observaciones
α	Nivel de significancia del test
β	Nivel de significancia dado por la desigualdad de Sidak
$\sigma_{y,i}$	Desvío estándar de la medición i-ésima
τ	Estadístico
χ^2	Distribución chi cuadrado
$\mathbf{\Sigma}$	Matriz de covarianza de las mediciones
$\mathcal{N}$	Distribución normal

4.9 Acrónimos

CM	Cuadrados Mínimos
DES	Detección de ESE
$\%DT_{ESE}$	Porcentaje de Detección Total de ESE
ECM	Error Cuadrático Medio
ECM_i	Error Cuadrático Medio de la i-ésima variable
ESE	Error Sistemático Esporádico
ET1	Error Tipo 1
$\%FA_{ESE}$	Porcentaje de Falsas Alarmas de ESE
MADN	Mediana normalizada de las desviaciones absolutas alrededor de la mediana
MSi	Método Simple
OP	Desempeño Global
RDC	Reconciliación de Datos Clásica
RDR	Reconciliación de Datos Robusta
RE	Redundancia Espacial
RE_i	Redundancia Espacial de la i-ésima variable
RT	Redundancia Temporal
TM	Test de las Mediciones
TMIM	Test de las Mediciones Iterativo Modificado
TRM	Test Robusto de las Mediciones
TRMV	Test de Razón de Máxima Verosimilitud
TMV	Test de las Mediciones de la Ventana

Capítulo 5

Tratamiento General de Errores Sistemáticos

5 Tratamiento General de Errores Sistemáticos

5.1 Introducción

La estrategia de Reconciliación de Datos Robusta (RDR) proporciona estimaciones insesgadas de las variables del proceso, que son consistentes con sus ecuaciones de balance, en presencia de una cantidad moderada de mediciones atípicas. Sin embargo, la presencia de Errores Sistemáticos que Persisten en el Tiempo (ESPT) puede ocasionar que se exceda el Punto de Quiebre (PQ) de las estimaciones, lo que perjudica el resultado de la RDR. Por tal motivo, en este capítulo se pondrá especial atención a la rápida detección e identificación de ESPT y al tratamiento de las mediciones afectadas con estos errores.

Las mediciones atípicas pueden ser Errores Sistemáticos Esporádicos (ESE) debidos a conexiones eléctricas deficientes, interferencias electromagnéticas, etc. Por el contrario, otras observaciones están acompañadas de errores debidos a la falta de calibración, ensuciamiento o deterioro de los sensores. Entre estos últimos son frecuentes los Sesgos (BI) y Derivas (DE). El BI es un error de magnitud constante en el tiempo, mientras que la DE presenta cierta funcionalidad con el mismo.

Numerosas investigaciones han demostrado las ventajas de usar M-estimadores robustos cuando las mediciones están contaminadas con valores atípicos. No obstante, sólo dos trabajos han tratado la identificación robusta de ESPT, y estos no proporcionaron análisis de desempeño adecuados (Martinez Prata y co., 2010; Zhang y Chen, 2015).

En este capítulo se presenta un algoritmo para la estimación robusta de las variables de un proceso y la clasificación de los errores sistemáticos en ESE, BI y DE (Llanos y

col., 2017). Este algoritmo hace uso de la metodología de reconciliación presentada en el Capítulo 3, denominada Método Simple, MSi; y del Test Robusto de las Mediciones (TRM), propuesto en el Capítulo 4. Asimismo, incorpora la técnica de Regresión Lineal Robusta (RLR) para categorizar los ESPT en BI o DE. Se analizan los resultados considerando métricas de desempeño relacionadas con la calidad de la estimación y las capacidades de detección e identificación de los distintos tipos de errores sistemáticos.

5.2 Motivación del Desarrollo

Las mediciones están sujetas a errores aleatorios que generan inconsistencias con las ecuaciones de conservación del proceso. Además de estos errores, las mediciones pueden estar contaminadas con errores sistemáticos que se presentan con baja frecuencia, tales como ESE, BI y DE (Narasimhan y Jordache, 2000; Bagajewicz, 2010). Esta contaminación puede originar que se exceda el PQ de las estimaciones obtenidas con el procedimiento de RDR. A continuación, se presenta un ejemplo numérico que ayuda a comprender el comportamiento de un M-estimador cuando los datos contienen diferentes errores sistemáticos.

5.2.1 Ejemplo numérico

Se formula el problema de RDR de la siguiente manera:

$$[\hat{x}_j^R,\ \hat{u}_j^R] = \underset{x_j,u_j}{Min} \quad \sum_{p=j-N+1}^{j} \sum_{i=1}^{I} \rho\left(\frac{y_{ip} - x_{ij}}{\sigma_{y,i}}\right)$$

$$st.$$

$$\begin{aligned} &\mathbf{f(x,u)=0} \\ &\mathbf{h(x,u)} \le \mathbf{0} \\ &\mathbf{x^L} \le \mathbf{x} \le \mathbf{x^U} \\ &\mathbf{u^L} \le \mathbf{u} \le \mathbf{u^U} \end{aligned} \quad , \qquad (5.1)$$

donde

$$a_{ij} = \frac{y_{ip} - x_{ij}}{\sigma_{y,i}}$$

es el ajuste estandarizado de la observación y_{ij}, $\sigma_{y,i}$ es el desvío estándar de la observación, ρ es la función de pérdida del M-estimador, N es la longitud de la ventana de datos, los vectores $[\hat{\mathbf{x}}_j^R, \hat{\mathbf{u}}_j^R]$ representan el estado del proceso en el j-ésimo intervalo de tiempo, y el modelo del proceso se compone de los sistemas de ecuaciones de igualdad, **f**, de desigualdad, **h**, y de los límites para las variables medidas **x** y no medidas **u**, cuyas dimensiones son I y U, respectivamente.

Se utilizan diferentes tipos de M-estimadores, cuyos parámetros se ajustan utilizando el método de Jackknife (Rey y co., 1983) de manera que sus Eficiencias Asintóticas (*Ef*) resulten iguales a 0,95.

Se consideran los siguientes modelos para la observación i-ésima (i=1…I)

- Modelo con errores aleatorios. Se supone que los errores aleatorios siguen la distribución normal estandarizada. El modelo se representa mediante la Ec. 5.2.

$$y_{ij} = x_i + \varepsilon_{ij}\,, \tag{5.2}$$

- Modelo con ESE. Con probabilidad p_G= 0,05, se agrega un ESE al modelo anterior de magnitud igual a $K_{ij}\sigma_{y,i}$. El número total de ESE generados es 3457.

$$y_{ij} = x_i + \varepsilon_{ij} + K_{ij}\sigma_{y,i}\,, \tag{5.3}$$

El parámetro K_{ij} indica la magnitud y signo del ESE en el tiempo j-ésimo, en este ejemplo se fija K_{ij}=10.

- Modelo con ESPT. Durante 100 intervalos de tiempo, se añaden ESPT al primer modelo. El número total de observaciones atípicas simuladas es 3600, que es una cantidad similar a la utilizada para el segundo modelo. La probabilidad de ocurrencia de BI y DE es la misma. La magnitud de los BI se fija en 6. La Ec. (5.4) se emplea para generar BI en las mediciones

$$y_{ij} = x_i + \varepsilon_{ij} + \mathrm{B}_{ij}\sigma_{y,i}\,, \tag{5.4}$$

donde B_{ij} es la magnitud del BI.

Además se considera que la DE tiene una variación lineal con el tiempo cuya pendiente es m_{drift}. A este parámetro se le asigna un valor igual a 1 en este ejemplo. El modelo empleado para simular DEs es el siguiente:

$$y_{ij} = x_i + \varepsilon_{ij} + m_{ij,drift}(t)\,\sigma_{y,i} = x_i + \varepsilon_{ij} + 1t\,\sigma_{y,i}\,. \tag{5.5}$$

Se utiliza un diagrama de flujo extraído del artículo de Rosenberg y co. (1987), que comprende 4 unidades y 7 corrientes medidas, es decir, el número de variables del problema de optimización es $I = 7$. El proceso se ilustra en la Fig. 5.1. Se considera que el desvío estándar de las mediciones es 2,5% respecto del valor verdadero. Se realizan 10.000 simulaciones del problema de RDR con $N = 20$. El Error Cuadrático Medio (ECM) se estima de la siguiente manera:

$$\mathrm{ECM}=\frac{1}{I\ Ns}\sum_{j=1}^{Ns}\sum_{i=1}^{I}\left(\frac{\hat{x}_{ij}^{R}-x_i}{\sigma_{y,i}}\right)^2, \tag{5.6}$$

siendo *Ns* el número de simulaciones. Los valores de ECM obtenidos usando los M-estimadores HU, CO, WE, BW y ERTP, y diferentes modelos de las observaciones se muestran en la Tabla 5.1.

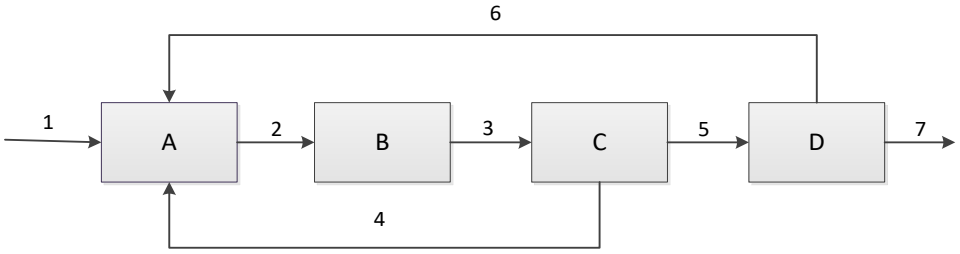


Figura 5.1: Diagrama de Flujo extraído de Rosenberg y co. (1987)

Tabla 5.1. ECM para diferentes M-estimadores y modelos de las observaciones

	ECM x 10^2		
M-estimadores	Aleatorios	ESE	ESPT
HU	3,114	3,596	51,394
CO	3,111	3,596	51,396
WE	3,118	3,596	51,406
BW	3,118	3,592	51,392
ERTP	3,118	3,595	51,392

Se observa que todos los M-estimadores minimizan el efecto perjudicial de los ESE y las estimaciones no son prácticamente afectadas por su presencia. No obstante, los ESPT degradan la estimación, lo cual se ve reflejado en el incremento de ECM. En consecuencia, los ESPT deben detectarse antes de que el PQ de las estimaciones obtenidas aplicando el procedimiento de RDR se exceda. Con este fin, se desarrollan estrategias dedicadas a detectar errores sistemáticos, clasificarlos y tomar las acciones correctivas apropiadas cuando se presentan ESPT en la Sección 5.4.

5.3 Concepto de Punto de Quiebre

El PQ de una estimación debe tenerse en cuenta cuando se aplican metodologías robustas. En términos generales, el punto de ruptura de una estimación $\hat{\vartheta}$ del parámetro ϑ es la mayor cantidad de contaminación (proporción de mediciones atípicas) que los datos pueden contener de tal manera que $\hat{\vartheta}$ todavía proporciona información cierta de ϑ.

Si se sabe que ϑ está contenida dentro de un conjunto Θ, y se quiere que $\hat{\vartheta}$ proporcionen información de ϑ, la contaminación no debe llevar a $\hat{\vartheta}$ al infinito o a los límites de Θ.

Definición 5.1 La contaminación asintótica, PQ, de una estimación $\hat{\vartheta}$ en *F*, denotada como $\xi^*\left(\hat{\vartheta}, F\right)$, es el máximo $\xi^* \in (0,1)$ tal que para $\xi < \xi^*$ se obtiene una estimación consistente, $\hat{\vartheta}_\infty$, $\hat{\vartheta}_\infty\left[(1-\xi)F + \xi G\right]$ como una función de *G*, que permanezca acotada y lejos de los límites del Θ.

Esta definición significa que existe un subconjunto $\mathbb{C} \subset \Theta$ tal que $\mathbb{C} \cap \Theta_{Límites} = \emptyset$, donde $\Theta_{límite}$ representa los límites del conjunto Θ, tal que:

$$\hat{\vartheta}_\infty\left[(1-\xi)F + \xi G\right] \in \mathbb{C} \;\forall\, \xi < \xi^* y \;\forall\, G \tag{5.7}$$

Es decir, existe un valor límite de contaminación por debajo del cual se pueden obtener estimaciones consistentes de ϑ. Se ha demostrado que este valor es $\xi^* \leq 1/2$ (Maronna y co., 2006) para estimaciones de localización equivariantes.

Para muestras finitas, el ξ^* puede calcularse como:

$$\xi^* \leq \frac{1}{n}\left[\frac{n-1}{2}\right] \tag{5.8}$$

donde *n* es número de mediciones utilizadas para el cálculo de $\hat{\vartheta}$. Esta aproximación es válida sólo para M-estimadores con función de influencia par y acotada. No existen desarrollos similares para M-estimadores redescendientes.

5.4 Nueva Estrategia de Reconciliación de Datos Robusta y Clasificación de Errores Sistemáticos

La estrategia propuesta interconecta tres procedimientos principales: RDR, TMR y RLR. Inicialmente se describen brevemente los mismos y luego se explican sus interconexiones.

5.4.1 Reconciliación de Datos Robusta

La metodología MSi se aplica para el ajuste de las mediciones en tiempo real. Esta metodología está compuesta por dos pasos que favorecen la combinación de las principales ventajas de los M-estimadores monótonos y redescendientes, y de la redundancia temporal provista por una ventana de mediciones.

Paso 1: estimación de la mediana robusta $\tilde{y}_i$

$$\tilde{y}_i = Min \sum_{p=j-N+1}^{j} \rho_{BW}\left(\frac{y_{ip} - \tilde{y}_i}{\hat{\sigma}_{y,i}}\right) \tag{5.9}$$

donde $\tilde{y}_i$ representa la mediana robusta de la ventana de mediciones y $\hat{\sigma}_y$ es la MAD de las mediciones.

Paso 2: estimación de $[\hat{\mathbf{x}}_j^R, \hat{\mathbf{u}}_j^R]$

$$[\hat{\mathbf{x}}_j^R, \hat{\mathbf{u}}_j^R] = \underset{x_j, u_j}{Min} \quad \sum_{i=1}^{I} \rho_{HU}\left(\frac{\tilde{y}_i - x_i}{\sigma_{y,i}}\right)$$

$$st.$$

$$\begin{aligned} &\mathbf{f(x,u)=0} \\ &\mathbf{h(x,u)} \le \mathbf{0} \\ &\mathbf{x^L} \le \mathbf{x} \le \mathbf{x^U} \\ &\mathbf{u^L} \le \mathbf{u} \le \mathbf{u^U} \end{aligned} \tag{5.10}$$

5.4.2. Test Robusto de las Mediciones

Para detectar mediciones atípicas se emplea el TRM. La prueba propuesta relaciona el vector de los ajustes de la medición, $\mathbf{a}_j^R = \mathbf{y}_j - \hat{\mathbf{x}}_j^{SIM}$, y su matriz de covarianza estimada $\hat{\mathbf{Q}}_j^R$. Para la variable i-ésima, el estadístico resulta igual a:

$$\hat{\tau}_{i,j}^R = \frac{\left|a_{ij}^R\right|}{\sqrt{\hat{Q}_j^R\big|_{ii}}} \tag{5.11}$$

Éste sigue una distribución de Student con un número de grados de libertad, *df = N*-1. El nivel de significancia de la prueba se establece en α = 0,025, con el cual se fija su valor crítico $\tau_{c_{\alpha,df}}^R$.

5.4.3 Regresión Lineal con CM

El análisis de regresión es una herramienta estadística para estimar la relación entre 2 o más variables. Consideremos la posibilidad de ajustar el siguiente modelo de regresión lineal

$$\hat{y} = \beta_0 + \beta_1 x \tag{5.12}$$

al conjunto de datos {(x_s, y_s): s = 1,… S}, donde x_s e y_s son los valores de las variables predictoras y de respuesta, respectivamente. El modelo de regresión puede presentarse en forma compacta como sigue:

$$\hat{y} = x_s \beta_s \tag{5.13}$$

donde $\beta_s = \begin{bmatrix} \beta_0 & \beta_1 \end{bmatrix}'$ y cada fila de matriz **X** está formada por el vector $\mathbf{x}_s = [1\ x_s]$ (s: 1 ... *S*).

El residuo puede calcularse como:

$$r_s = y_s - \hat{y} \tag{5.14}$$

$$r_s = y_s - x_s \beta_s \tag{5.15}$$

La estimación de β es el $\hat{\beta}_{CM}$ tal que:

$$Min \sum_{s=1}^{S} r_s^2 \left(\hat{\beta}_{CM} \right) \tag{5.16}$$

Derivando está expresión resulta:

$$\sum_{s=1}^{S} r_s \left(\hat{\beta}_{CM} \right) \mathbf{x}_s = \mathbf{0} \tag{5.17}$$

De forma vectorial r puede escribir como:

$$\mathbf{r} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}_{CM} \tag{5.18}$$

Y el equivalente de la ecuación 5.15 es

$$\mathbf{X}^T \left(\mathbf{X}\hat{\boldsymbol{\beta}}_{CM} - \mathbf{y} \right) = \mathbf{0} \tag{5.19}$$

Si **X** es una matriz de rango completo

$$\hat{\boldsymbol{\beta}}_{CM} = \left(\mathbf{X}^T \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{y} \tag{5.20}$$

Asumiendo que los elementos de $\mathbf{y}_s$ son mediciones independientes e idénticamente distribuidas y siguen una distribución normal con varianza **Σ** y, considerando que el modelo es lineal

$$E\left(\hat{\boldsymbol{\beta}}_{CM} \right) = \boldsymbol{\beta} \tag{5.21}$$

$$\text{var}(\hat{\boldsymbol{\beta}}_{CM}) = \Sigma\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \tag{5.22}$$

entonces:

$$\hat{\boldsymbol{\beta}}_{CM} \sim N_p\left(\boldsymbol{\beta}, \Sigma\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\right) \tag{5.23}$$

5.4.4 Regresión Lineal Robusta

Los métodos de RLR están basados en el principio de Máxima Verosimilitud (Maronna y co., 2006). Tienen como objetivo dar un buen ajuste a la mayor parte de los datos sin ser perturbados por una pequeña proporción de mediciones atípicas.

El vector de la regresión cuando se emplean M-estimadores, $\hat{\beta}_s^R = [\hat{\beta}_0^R \ \hat{\beta}_1^R]$ se define como la solución del siguiente problema de optimización:

$$Min \quad \sum_{s=1}^{S} \rho\left(\frac{r^R{}_s\left(\hat{\beta}_s^R\right)}{\hat{\sigma}_r}\right), \tag{5.24}$$

donde

$$r^R{}_s = y_s - \hat{y}_s = y_s - \left(\hat{\beta}_0^R + \hat{\beta}_1^R x_s\right), \tag{5.25}$$

$$r^R{}_s = y_s - x_s\hat{\beta}^R \tag{5.26}$$

y $\hat{\sigma}_r$ es su estimación de escala, que puede calcularse como:

$$\hat{\sigma}_r = \frac{1}{0.675} Med\left(\left|r^R{}_s\right| \quad \left|r^R{}_s \neq 0\right.\right) \tag{5.27}$$

La condición necesaria y suficiente para resolver el problema (5.9) es:

$$\sum_{s=1}^{S} \psi\left(\frac{\hat{r}_s^R}{\hat{\sigma}_r}\right)\mathbf{x}_s = 0 \tag{5.28}$$

Usando la ecuación (3.10), Ψ se puede sustituir por la función de peso *W* como sigue

$$\sum_{s=1}^{S} W\left(\frac{\hat{r}_s^R}{\hat{\sigma}_r}\right)\hat{r}_s^R\,\mathbf{x}_s = \mathbf{0} \tag{5.29}$$

Reemplazando $\hat{r}_s^R$ se obtiene

$$\sum_{s=1}^{S} W\left(\frac{\hat{r}_s^R}{\hat{\sigma}_r}\right)\left(y_s - \mathbf{x}_s\boldsymbol{\beta}^R\right)\mathbf{x}_s = \mathbf{0} \tag{5.30}$$

$$\begin{bmatrix} \sum_{s=1}^{S} W\left(\frac{\hat{r}_s^R}{\hat{\sigma}_r}\right)\left(y_s - \left(\hat{\beta}_0^R + \hat{\beta}_1^R x_s\right)\right) \\ \sum_{s=1}^{S} W\left(\frac{\hat{r}_s^R}{\hat{\sigma}_r}\right)\left(y_s x_s - \left(\hat{\beta}_0^R + \hat{\beta}_1^R x_s\right)x_s\right) \end{bmatrix} = \begin{bmatrix} 0 \\ \\ 0 \end{bmatrix} \tag{5.31}$$

De la ecuación (5.31), surgen las siguientes expresiones para estimar los parámetros del modelo de regresión:

$$\hat{\beta}_0^R = \frac{\sum_{s=1}^{S}\left[W\left(\frac{\hat{r}_s^R}{\hat{\sigma}_r}\right)\left(y_s - \hat{\beta}_1^R x_s\right)\right]}{\sum_{s=1}^{S}\left[W\left(\frac{\hat{r}_s^R}{\hat{\sigma}_r}\right)\right]} \tag{5.32}$$

$$\hat{\beta}_1^R = \frac{\sum_{s=1}^{S}\left[W\left(\frac{\hat{r}_s^R}{\hat{\sigma}_r}\right)x_s\left(y_s - \hat{\beta}_0^R\right)\right]}{\sum_{s=1}^{S}\left[W\left(\frac{\hat{r}_s^R}{\hat{\sigma}_r}\right)x_s^2\right]} \tag{5.33}$$

La varianza de $\hat{\boldsymbol{\beta}}_s^R$ se calcula como:

$$\text{var}\left(\hat{\boldsymbol{\beta}}_s^R\right) = \hat{\mathbf{v}}_s\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \tag{5.34}$$

donde:

$$\hat{\mathbf{v}}_s = \hat{\sigma}^2\,\frac{ave\left[\psi\left(\frac{r^R{}_s}{\hat{\sigma}_r}\right)^2\right]}{\left(ave\left[\psi'\left(\frac{r^R{}_s}{\hat{\sigma}_r}\right)\right]\right)^2}\,\frac{S}{S-2} \tag{5.35}$$

5.4.5 Test de la Pendiente

El resultado de RLR se utiliza para calcular un estadístico que permite clasificar un ESPT. Se plantean las siguientes hipótesis estadísticas:

H_0: $\hat{\beta}_1^R = 0$

H_1: $\hat{\beta}_1^R \neq 0$

Y el estadístico $T_{\beta 1}$, que considera la relación entre $\hat{\beta}_1$ y su varianza, se define como:

$$T_{\beta_1} = \frac{\hat{\beta}_1^R}{\left[\text{var}\left(\hat{\beta}_1^R\right)\right]^{1/2}} \sim t_{df} \tag{5.36}$$

y sigue la distribución Student con *df* = *S*-2. El valor del $T_{\beta 1,c}$ queda definido por *df* y α = 0,05. Cuando $T_{\beta 1} < T_{\beta 1,c}$, se satisface la H_0 y se dice que el ESPT es un BI, en caso contrario el error se clasifica como DE. En Matlab los parámetros del modelo $\left[\hat{\beta}_0^R \quad \hat{\beta}_1^R\right]$ se calculan iterativamente utilizando como valor inicial la solución provista por la regresión lineal con CM. A continuación se explica el algoritmo que interconecta los procedimientos RDR, TRM y RLR.

5.4.6 Algoritmo del Método Propuesto

Para cada tiempo *j* la estrategia consta de tres etapas. En la primera, se resuelve el problema de RDR. La segunda etapa está relacionada con la clasificación de los errores, y la tercera, en caso de que sea necesario, actualiza las entradas al siguiente problema de RDR.

1) Primera Etapa: La probabilidad experimental, basada en 10.000 observaciones, de que cuatro $\hat{\tau}_i^R$ consecutivos superen el valor del estadístico crítico para α = 0,025 es cero para cualquier *N* cuando las mediciones siguen exactamente la distribución normal. Este hecho se utiliza para distinguir entre ESE y ESPT, pues si el TMR detecta este evento

para la *i*-ésima variable, el sensor se considera defectuoso. Sus mediciones son reemplazadas durante los siguientes intervalos de tiempo por los valores generados usando la solución de la RDR obtenida para la última observación antes que la variable presentase problemas. Esto evita la contaminación de la ventana de datos con observaciones atípicas mientras se ejecutan otras acciones correctivas.

2) Segunda Etapa: Cuando se dispone de *N*/2 mediciones atípicas de la *i*-ésima variable, se aplica la RLR para categorizar el ESPT previamente detectado como un BI o una DE y se informa el problema al grupo de mantenimiento. En el caso en que el ESPT es clasificado como BI, su magnitud se estima cuando se toman *N* mediciones atípicas. Los siguientes problemas de estimación utilizan la medición corregida a partir del BI calculado hasta que el sensor se repara.

3) Tercera Etapa: Actualización de las entradas a la RDR. Se define un vector $\boldsymbol{s}$ de sensores defectuosos que incluye los instrumentos para los que se ha detectado un ESPT. En la primera etapa, el nuevo vector de observación $\mathbf{y}_j$ se incorpora a la matriz de medición $\mathbf{Y_{ob}}$ (*I*, *N*) eliminando el vector de observaciones más antiguo de la matriz y añadiendo $\mathbf{y}_j$ como última columna. Si $\boldsymbol{s}$ = [], esa matriz se utiliza como dato de entrada del problema de RDR. Si no es así, $\mathbf{Y_{ob}}$ se modifica como se explicará más adelante y se obtiene la matriz de medición $\mathbf{Y_{ob}}$*. Luego se ejecuta el MSi para obtener $\hat{\mathbf{x}}_j^{SIM}$.

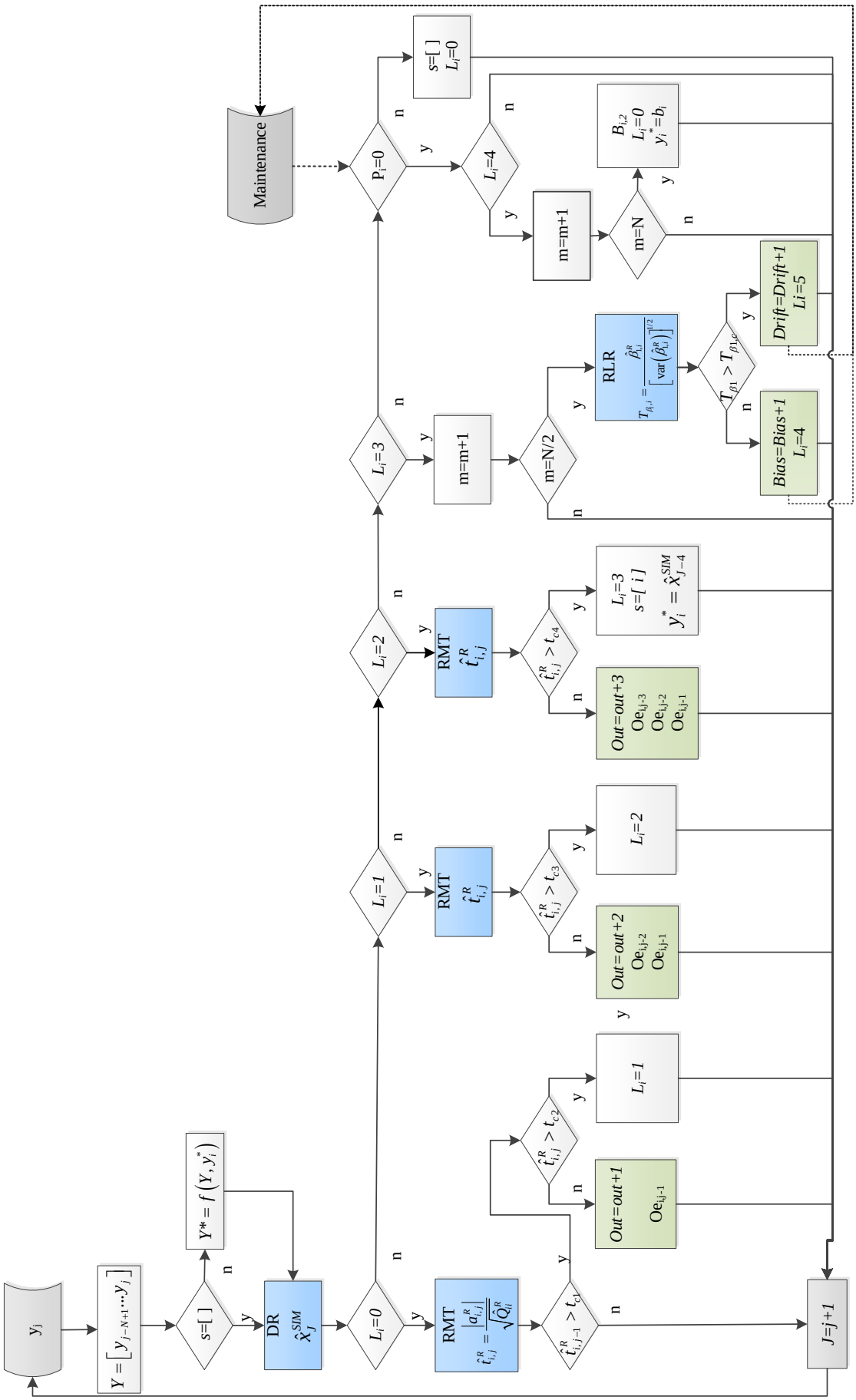


Figura 5.2 Diagrama de flujo de la metodología propuesta

La Fig. 5.2 es el diagrama de flujo de la estrategia propuesta; en ésta se observa que las tareas involucradas en la segunda y tercera etapas dependen de los valores de los índices η_i y ϖ_i (i = 1,..., I). El primero resume los resultados obtenidos durante los intervalos de tiempo anteriores con respecto a la clasificación de un error sistemático para la *i*-ésima variable. El segundo es una variable binaria que indica si el sensor ha sido reparado (ϖ_i = 1) o no (ϖ_i = 0). En la Tabla 5.2, se muestran los valores iniciales de las variables del algoritmo.

Tabla 5.2 Valores iniciales de las variables

Variable	Valor
s	[ ]
η_i	0
ϖ_i	1

El índice η_i puede variar en el rango [0 1 2 3 4 5] y su valor inicial es η_i = 0. El estadístico $\tau_{i,j}^R$ se evalúa después de la etapa de reconciliación. Sólo cuando η_i = 0 se calculan dos estadísticos el del tiempo actual y el del anterior. A continuación se explica las posibles variaciones de este índice.

- Si η_i = 0, se detecta un ESE para la penúltima medida si $\tau_{i,j-1}^R$ es mayor que el valor crítico pero el $\tau_{i,j}^R$ de la última medición no excede este límite. Por el contrario, si ambos $\tau_{i,j-1}^R$ y $\tau_{i,j}^R$ superan el valor crítico, no hay suficiente información para clasificar el conjunto de dos observaciones con ESE, y η_i se establece igual a 1. Debido a que

la presencia de una o dos observaciones atípicas no afecta los resultados de la RDR para los valores de *N* comúnmente utilizados, se puede continuar sin cambiar la matriz $\mathbf{Y_{ob}}$;

- Si η_i = 1 y $\tau_{i,j}^R$ no es mayor que el valor crítico, se identifica la presencia de dos mediciones atípicas consecutivas, luego de esto el índice η_i se restablece en 0, en caso contrario η_i =2;
- Si η_i = 2 y $\tau_{i,j}^R$ no es mayor que el valor crítico, se identifica la presencia de tres mediciones atípicas consecutivos, luego de esto el índice η_i se restablece en 0. Por el contrario, se considera que la secuencia de cuatro observaciones inusuales es parte de un ESPT. En este caso, la *i*-ésima variable se incluye en el vector $\boldsymbol{s}$ y η_i se fija igual a 3. También se generan *N* observaciones con errores aleatorios usando el valor ajustado de la *i*-ésima variable obtenido antes del comienzo del ESPT ($\hat{\mathbf{x}}_{i,j-4}^{SIM}$). Éstas formarán la *i*-ésima fila de la matriz $\mathbf{Y_{ob}}$* utilizada para la siguiente ejecución del procedimiento de RDR;
- Si η_i = 3, las nuevas mediciones de la variable *i*-ésima no se utilizan para la RDR, y no tiene sentido calcular su estadístico. Estas observaciones se guardan hasta que se obtienen *N*/2 mediciones consecutivas atípicas. En ese momento, se ejecuta la técnica RLR, y la estimación de la pendiente de la recta se utiliza para decidir si el error sistemático es un BI (η_i =4) o una DE (η_i = 5). Se envía una alerta al grupo de mantenimiento de instrumentación cambiando el valor de la variable binaria ϖ_i a 0;
- Si η_i = 4, las nuevas medidas se guardan hasta que está disponible una ventana completa de observaciones contaminadas por este error. En este momento, una nueva

estimación de BI, B_{i2}, se calcula como la diferencia entre la mediana robusta de las mediciones y el valor reconciliado de la variable, entonces $\eta_i = 0$. El vector fila $\mathbf{b_i}$ de dimensión *N*, cuyos elementos son iguales a B_{i2}, se envía como entrada del problema RDR para corregir la *i*-ésima fila de $\mathbf{Y_{ob,j+1}}$ como sigue:

$$\mathbf{Y}^{*}_{ob,j+1}(i,:) = \mathbf{Y}_{ob,j+1}(i,:) - \mathbf{b_i} \tag{5.37}$$

hasta que el sensor haya sido reparado. Cuando esto sucede, $\varpi_i = 1$, entonces **s** = [], y $\eta_i = 0$;

- Si $\eta_i = 5$, no se pueden realizar otras tareas si se ha identificado una DE porque las mediciones pueden seguir diversos comportamientos en el tiempo. Las mediciones de esta variable son reemplazadas por las estimaciones de la RDR anterior hasta que el sensor haya sido reparado. Cuando esto sucede, $\varpi_i = 1$, entonces **s** = [], y $\eta_i = 0$.

El método propuesto permite estimar la magnitud del ESE como la diferencia entre la medida y el valor de la variable ajustada obtenida usando el procedimiento de RDR. La estimación del BI se realiza como se ha explicado anteriormente. En cuanto a la DE, si se supone que la ésta puede representarse por una función lineal del tiempo y se estima su pendiente, entonces es posible calcular la evolución temporal de la magnitud del error. En cambio si $m_{drift}(t)$ es desconocida, la estimación del error no puede realizarse.

En la Figura 5.2 se usan las variables *out*, *bias* y *drift* para cuantificar la clasificación de los diferentes errores sistemáticos. Además, se informa la variable y el tiempo en que el algoritmo realiza la clasificación. Por ejemplo, $O_{i,\,j-1}$ se refiere a la presencia de un outlier en la *i*-ésima variable en el (*j*-1) -ésimo intervalo de tiempo.

La Fig. 5.3 ilustra una secuencia temporal de vectores de observación y muestra cómo se distingue la presencia de un ESE o de ESE consecutivos de la ocurrencia de un ESPT. Los estadísticos cuyos valores son más pequeños que el crítico están representados por un punto, de lo contrario se simbolizan usando ⊗.

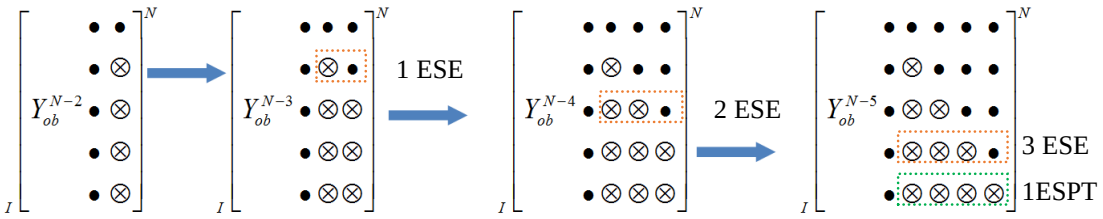


Figura 5.3 Secuencia de errores sistemáticos

5.5 Análisis de desempeño

Se realiza un análisis de desempeño de la metodología desarrollada para cuantificar sus capacidades de detección y clasificación de los errores sistemáticos presentes en las mediciones de dos procesos químicos. Estos son: la Red de Ingreso de Vapor (SMN) (Serth y Hennan, 1986) y la Red de Intercambiadores de Calor (HEN) (Swartz, 1989), los cuales se presentan frecuentemente en la literatura sobre RDR. Para cada proceso, se proponen cuatro casos de estudios que implican fijar diferentes valores de los parámetros N, B y m_{drift}, para una magnitud fija de K=10. La Tabla 5.3 presenta los valores de los parámetros e indica si se ha aplicado o no la metodología propuesta para el correspondiente caso de estudio.

El Caso I muestra los resultados de la aplicación de la estrategia cuando se simulan sólo errores aleatorios. En cambio, el Caso II presenta su comportamiento cuando las mediciones están contaminadas con ESE, BIs y DEs. El Caso III es la peor condición para probar la metodología porque las magnitudes de los BIs y las DEs son menores que las

establecidas para el Caso II. En los Caso IV y V, se muestran los resultados cuando no se aplica la estrategia propuesta y las magnitudes de los errores son las de los Casos II y III, respectivamente.

Tabla 5.3 Descripción de los casos de estudio

Caso	K	B	m_{drift}	Metodología
I	0	0	0	Sí
II	10	6	1	Sí
III	10	4,5	0,75	Sí
IV	10	6	1	No
V	10	4,5	0,75	No

Se considera que *Ns*=10.000 para cada caso de estudio. Las observaciones simuladas contienen errores aleatorios (Ec. 5.2) y se contaminan aleatoriamente con errores sistemáticos. La probabilidad de ocurrencia de errores sistemáticos en una medición es *p* = 0,02. Éste procedimiento aleatorio permite simular la presencia de ESE consecutivos, así como su aparición simultánea con BI o DE. El 95% de los errores sistemáticos simulados son ESE, y el resto son BI y DE en igual proporción. Se considera que estos últimos persisten durante 100 intervalos de tiempo, por lo tanto, aproximadamente el 12% de las observaciones simuladas son atípicas. Esto es:

$$p_G = \frac{0,02 Ns\, I(0,95x1 + 0,05x100)}{Ns\ \ I}$$

Una vez que se detecta un ESPT, la acción correctiva se realiza sobre 100 intervalos de tiempo, tras lo cual se asume que el sensor está disponible nuevamente. Se considera que un instrumento que experimentó un ESPT, no tendrá un BI o DE en un lapso de 400 intervalos después que el error haya concluido. Ésto indica que un equipo reparado no

vuelve a quedar fuera de servicio de inmediato, sino que existe un período en el cual el instrumento funcionará correctamente.

El procedimiento se ejecuta con un procesador Intel® Core ™ i7 CPU 930 2,80 GHz, 8GB de RAM; se utiliza el código de Programación Cuadrática Sucesiva de MatLab Release 7.12 (R2011a) para resolver el problema de optimización no lineal.

Se proponen índices de desempeño globales e individuales. Inicialmente, se definen parámetros que permiten seguir el comportamiento de la estrategia cuando una variable se declara sospechosa. Estos son la capacidad de detección total de ESPT (% DT_{ESPT}) y el Porcentaje de Falsas Alarmas de ESPT (% FA_{ESPT}), los cuales se formulan a continuación:

$$\%\mathrm{DT}_{ESPT} = \frac{(\#ESPT)_{Simulados\ y\ Detectados}}{(\#ESPT)_{Simulados}} x\,100 \tag{5.38}$$

$$\%\mathrm{FA}_{ESPT} = \frac{(\#ESPT)_{Detectados} - (\#ESPT)_{Simulados\ y\ CorrectamenteDetectados}}{(\#ESPT)_{Detectados}} x\,100 \tag{5.39}$$

Las detecciones erróneas de ESPT causan falsas alarmas, pues se emite una señal al grupo de mantenimiento cada vez que se detecta un ESPT. Estos eventos se cuantifican utilizando el $\%FA_{ESPT}$. Dado que la ocurrencia de una cantidad moderada de ESE no afecta significativamente la exactitud de las estimaciones si se aplican procedimientos de RDR, las falsas alarmas debidas a la aparición de ESE no se evalúan.

Asimismo, se utilizan parámetros de desempeño global como el ECM, Ec. (5.6), y el porcentaje de Detección Total de Errores Sistemáticos (% DT). Este último representa

el porcentaje de errores sistemáticos simulados que se detectan, y se calcula de la siguiente forma:

$$\%\text{DT}=\frac{\left(\#ESE+\#ESPT\right)_{Simulados\ y\ Detectados}}{\left(\#ESE+\#ESPT\right)_{Simulados}}\,x\,100 \tag{5.40}$$

Como se mostró en la Sección 5.2, el ECM tiene en cuenta la diferencia entre los valores de las variables reconciliadas y los valores verdaderos. A pesar de que un estimador robusto es capaz de reducir el efecto de una cantidad moderada de observaciones atípicas, la presencia de un ESPT sesga los resultados del problema RDR y provoca el incremento del ECM. Por lo tanto, es útil analizar la influencia global del procedimiento de detección y clasificación de ESPT en los valores de las variables reconciliadas.

Los índices individuales de desempeño para los diferentes tipos de ESPT se evalúan en términos de sus porcentajes de detección y clasificación correcta y errónea. Para los ESE, sólo se calcula un índice pues la detección e identificación se realizan en un único paso. La Fig. 5.4 muestra la relación entre las diferentes medidas de desempeño.

Los errores sistemáticos simulados se dividen en Detectados y No Detectados teniendo en cuenta los resultados del TRM. Si se detectan errores aislados, se clasifican directamente como ESE. Por el contrario, si la variable es sospechosa el procedimiento de RLR se aplica para categorizarla como un BI o DE.

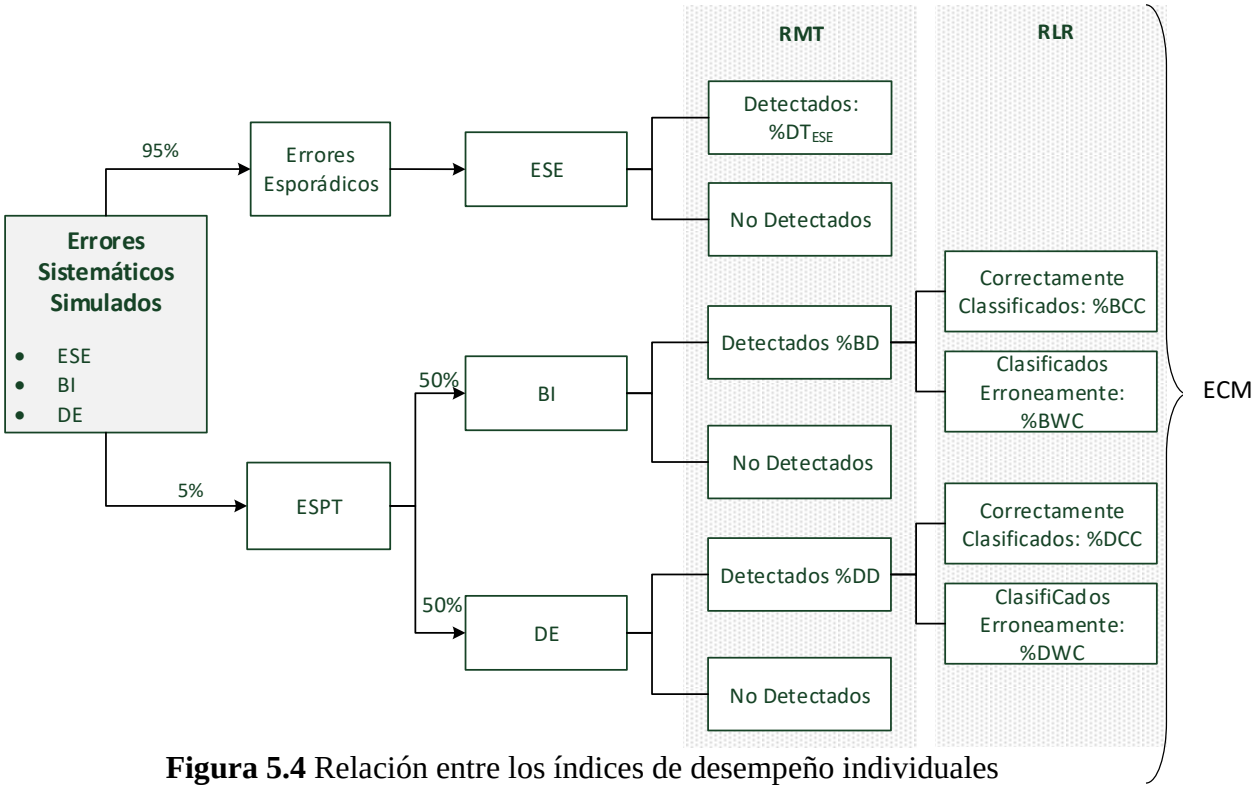


Figura 5.4 Relación entre los índices de desempeño individuales

Con respecto a los BIs, los detectados forman parte del Porcentaje de BI Detectado (% BD), que se divide en el Porcentaje de BI Correctamente Clasificados (% BCC) y el Porcentaje de BIs Clasificados Erróneamente (% BCE). Se definen las mismas medidas de desempeño para las DEs. El Porcentaje de DEs Detectadas (% DD) se descompone en el Porcentaje de DEs Correctamente Clasificadas (% DCC) y el Porcentaje de DEs Clasificadas Erróneamente (% DCE). Es decir:

$$\%\mathrm{DT}_{ESE} = \frac{(\#\,ESE)_{\,SimuladosyCorrectamenteClasificados}}{(\#\,ESE)_{\,Simulados}}\,x\,100 \tag{5.41}$$

$$\%\mathrm{BD} = \frac{(\#\,BI)_{\,Simulados\ y\ Correctamente\ Detectados}}{(\#\,BI)_{\,Simulados}}\,x\,100 \tag{5.42}$$

$$\%\mathrm{BCC} = \frac{(\#\,BI)_{\,Correctamente\ Clasificado}}{(\#\,BI)_{\,Simulados}}\,x\,100 \tag{5.43}$$

$$\%\text{BCE}=\frac{(\#BI)_{Mal\ Clasificado}}{(\#BI)_{Simulados}}\,x\,100 \tag{5.44}$$

$$\%\text{DD}=\frac{(\#DE)_{Simuladas\ y\ Correctamente\ Detectadas}}{(\#DE)_{Simuladas}}\,x\,100 \tag{5.45}$$

$$\%\text{DCC}=\frac{(\#DE)_{Correctamente\ Clasifcadas}}{(\#DE)_{Simuladas}}\,x\,100 \tag{5.46}$$

$$\%\text{DCE}=\frac{(\#DE)_{Mal\ Clasificada}}{(\#DE)_{Simuadas}}\,x\,100 \tag{5.47}$$

Para cada conjunto de 10.000 simulaciones se almacena el tiempo de detección (θ_d) de los BI y DE, y se calcula: la mediana de los tiempos de detección, $\tilde{\theta}_d$, el número de veces que $\theta_d > \tilde{\theta}_d$, *Nv*, y la media de los valores de θ_d para los cuales se verifica que $\theta_d > \tilde{\theta}_d$, $\bar{\theta}_{d,Nv}$. Estas medidas permiten analizar los casos en los cuales θ_d supera *N*/2 y la primera etapa de la metodología MSi provee estimaciones sesgadas a la etapa de RDR.

Para probar el desempeño de un método que detecta observaciones contaminadas con errores sistemáticos, ha sido una práctica común generar vectores de observaciones con dichos errores y ejecutar la estrategia. Si se emplea dicha práctica, el resultado de cada ensayo de simulación no está relacionado con el resultado del ensayo anterior. Sin embargo, la metodología que se propone en este capítulo realiza diferentes acciones dependiendo de la evolución temporal de cada medición. Por lo tanto, resultó necesario desarrollar el procedimiento de prueba *ad hoc* descripto para la nueva estrategia desarrollada.

5.6 Análisis de los Resultados

La metodología propuesta se aplicó a dos procesos químicos y se obtuvieron los resultados que se detallan a continuación.

5.6.1 Red de Ingreso de Vapor (SMN)

El SMN involucra 28 corrientes que interconectan 11 unidades. Los caudales de todas las corrientes se consideran medidos. Se generan errores aleatorios asumiendo que el desvío estándar de las observaciones es 2,5% de sus valores verdaderos.

Las medidas de desempeño se calculan para los Casos II y III, excepto el ECM que se evalúa para todos los casos de estudio. Para el Caso II, los índices globales e individuales se presentan en las Tablas 5.4 y 5.5, respectivamente, y la Tabla 5.6 muestra los valores de $\hat{\theta}_d$, *Nv* y $\bar{\theta}_{d,Nv}$ para los ESPT. La misma información se presenta en las Tablas 5.7, 5.8 y 5.9 para el Caso III. En la Tabla 5.10 se muestran los valores del ECM.

Se generan 228 ESPT en diferentes variables, que permanecen durante 100 intervalos de tiempo. Los límites de *N* considerados se fijan considerando una capacidad de detección mínima del 75% de los ESPT simulados y el deterioro del ECM con *N*.

Tabla 5.4. Índices de Desempeño Global vs *N* - Caso II (SMN)

N	% DT_{ESPT}	% FA_{ESPT}	%DT	ECM
24	78,07	18,72	86,10	297,079
26	95,18	3,98	97,23	8,425
28	96,05	11,34	96,14	20,507
30	99,56	0,44	98,19	0,116
40	100	0,870	97,63	0,140

Tabla 5.5 Índices de Desempeño Individual vs *N* - Caso II (SMN)

N	% DT_{ESE}	%BD	%BCC	%BCE	%DD	%DCC	%DCE
24	86,42	82,69	77,88	4,81	74,19	72,58	1,61
26	97,31	94,23	86,54	7,69	95,97	95,16	0,81
28	96,14	94,23	90,38	3,85	97,58	97,58	0
30	98,13	99,04	94,23	4,81	100	100	0
40	97,54	100	94,23	5,77	100	100	0

Tabla 5.6. Tiempos de Detección vs *N* - Caso II (SMN)

N	θ_d ,BCC			θ_d ,BCE			θ_d ,DCC			θ_d ,DCE		
	$\hat{\theta}$	*Nv*	$\bar{\theta}_{d,Nv}$	$\hat{\theta}$	*Nv*	$\bar{\theta}_{d,Nv}$	$\hat{\theta}$	Nv	$\bar{\theta}_{d,Nv}$	$\hat{\theta}$	*Nv*	$\bar{\theta}_{d,Nv}$
24	4	4	75,5	4			7	26	8,81	55	2	55
26	4	1	68	4	1	6	7	34	8,79	46	1	46
28	4	2	54	4			7	31	8,65			
30	4	1	7	4			7	32	8,63			
40	4			4			6	58	7,55			

Con respecto al Caso II, se observa que a mayor *N* los índices individuales de detección y clasificación de BI y DE se ven favorecidos. Se destaca que para N≥30 los ESPT son detectados antes de N/2, con lo cual se provee de un buen punto inicial a la RDR. Además se distinguen los siguientes comportamientos con *N*:

N=24: Se obtiene el menor desempeño del método de todas las pruebas realizadas, esto se atribuye al bajo %DT_{ESPT}. Sobre los ESPT no detectados no se aplican metodologías correctivas, en consecuencia estas mediciones provocan el sesgamiento de los resultados de la RDR. Esto perjudica el desempeño del TRM disminuyendo su capacidad de detección y generando falsas alarmas. Además, con este *N* se detectan variables de forma tardía. Todos estos factores afectan al ECM, el cual alcanza el máximo para este caso de estudio.

N=26: El $\%DT_{ESPT}$ se incrementa un 10% respecto de la ventana precedente y el ECM disminuye. Los parámetros de desempeño global e individual se ven favorecidos con la longitud de esta nueva ventana.

N=28: El $\%DT_{ESPT}$ incrementa menos de un 1% respecto de la ventana precedente y el ECM aumenta. En esta prueba se observa que la principal diferencia, respecto de la ventana anterior, se encuentra en el incremento de $\%FA_{ESPT}$. No obstante, esta experiencia es la única que no sigue la tendencia general y será objeto de futuros estudios.

N=30: El $\%DT_{ESPT}$ incrementa y el ECM disminuye. Con esta longitud de ventana se observa una mejora en los parámetros de desempeño global e individual. A partir de este *N*, todas las DEs son correctamente clasificadas y todos los ESPT son detectados en sus inicios.

N=40: Se alcanza el 100% de DT_{ESPT}, sin embargo, el ECM aumenta respecto de la ventana anterior. El $\%DT_{ESE}$ y %DT se ven afectados por el incremento de falsas alarmas. Éstas provocan que se ejecutan acciones correctivas que perjudican a la RDR y hacen aumentar el ECM.

Se observa que el $\%FA_{ESPT}$ afecta al desempeño de la metodología, las falsas alarmas se presentan cuando:

- Los ESPT no son detectados. Un ESPT simulado, persiste en la variable durante 100 mediciones, por lo tanto, si éste no es detectado al inicio provoca un sesgamiento en el valor reconciliado de la variable. Al finalizar las simulaciones con dicho ESPT, se reciben nuevas mediciones con error aleatorio las cuales distan del valor reconciliado. Esta situación provoca que el TRM entregue τ mayores al crítico y dan lugar a falsas alarmas.

- Disminución del τ_c. El aumento de *N* provoca un leve aumento en las falsas alarmas, esto se debe a que el τ_c del TMR disminuye y por lo tanto aumenta el ET1.

Asimismo, el aumento de %FA provoca:

- Incrementos en el ECM porque se ejecutan acciones correctivas innecesarias que introducen error en las mediciones.
- Disminución en $\%DT_{ESE}$, pues el TRM no se ejecuta cuando una variable es clasificada como sospechosa.
- Disminución en %DT. El 95% de los errores sistemáticos simulados son ESE, es por esto que el %DT está más influenciado por el $\%DT_{ESE}$ que por el $\%DT_{ESPT}$, esto provoca que se vea deteriorado en el mismo sentido que la detección de ESE

Estas observaciones se aplican a todos los casos de estudio. A continuación, se presentan los resultados del Caso III

Tabla 5.7. Índices de Desempeño Global vs *N* - Caso III (SMN)

N	$\%DT_{ESPT}$	$\%FA_{ESPT}$	%DT	ECM
30	78,07	8,25	97,48	6,282
40	92,54	6,64	97,55	0,218
50	96,930	0,897	97,631	0,129
60	99,123	0,441	97,076	0,134

Tabla 5.8. Índices de Desempeño Individual vs *N* - Caso III (SMN)

N	% DT_{ESE}	%BD	%BCC	%BCE	%DD	%DCC	%DCE
30	98,25	64,42	60,58	3,85	89,52	89,52	0
40	97,75	83,65	79,81	3,85	100	100	0
50	97,66	93,27	85,58	7,69	100	100	0
60	96,995	98,077	93,269	4,808	100	100	0

Tabla 5.9. Tiempos de Detección vs *N* - Caso III (SMN)

N	θ_d ,BCC			θ_d ,BCE			θ_d DCC			θ_d DCE		
	$\hat{\theta}$	*Nv*	$\bar{\theta}_{d,Nv}$	$\hat{\theta}$	*Nv*	$\bar{\theta}_{d,Nv}$	$\hat{\theta}$	*Nv*	$\bar{\theta}_{d,Nv}$	$\hat{\theta}$	*Nv*	$\bar{\theta}_{d,Nv}$
30	4	2	44	4	1	6	8	39	10,18			
40	4	5	6,6	4	1	6	8	38	9,66			
50	4	10	7,1	4	1	6	7	55	8,62			
60	4	9	6,7	4	1	9	7	49	8,51			

Para el Caso III, se observa que los índices globales en general siguen las mismas tendencias que las presentadas por el Caso II, es decir el $\%DT_{ESPT}$ aumenta con *N* y lo contrario ocurre con ECM. Esto sucede hasta alcanzar un *N* en el que el ECM se deteriora (*N*=60) por esto luego se comparan los índices de esta ventana con los correspondientes a *N*=50.

Algunas consideraciones sobre el comportamiento de los índices individuales son:

- En general las medidas de desempeño y $\hat{\theta}$ se ven favorecidas con el aumento de *N*, con excepción del $\%DT_{ESE}$.
- Las DEs detectadas se clasificadan correctamente. Además, cuando $N \geq 40$ se alcanza el 100% de DD.

- La comparación entre *N*=50 y *N*=60 muestra que todos los índices individuales mejoran con la longitud de ventana. Además, se observa que el %BCE es mayor en *N*=50, con lo cual, una mayor cantidad de BIs son tratados como DEs y por lo tanto la RDR da resultados más exactos.

Las observaciones con ESPT clasificados como DEs, son reemplazadas por valores obtenidos a partir de estimaciones previas. Por otro lado, las mediciones con BIs son corregidas con la estimación de la magnitud del sesgo. Este último tratamiento es menos preciso que el efectuado a las mediciones con DE y por lo tanto introduce error, que repercute sobre la RDR. Asimismo, a mayor *N* aumenta la probabilidad de tener múltiples errores sistemáticos en las variables, lo cual perjudica a la RDR. Estas dos situaciones pueden estar presentándose en *N*=60 y afectando el ECM.

Los ECM de los 5 casos considerados se presentan en la Tabla 5.10

Tabla 5.10. ECM vs *N* (SMN)

N	Caso I	Caso II	Caso III	Caso IV	Caso V
24	0,027	297,079	--	3.494,363	--
26	0,025	8,425	--	3.069,751	--
28	0,023	20,507	--	2.921,332	--
30	0,022	0,116	6,282	3.189,345	1.709,73
40	0,016	0,140	0,218	3.424,207	1.248,48
50	0,013	0,172	0,129	2.279,945	1.465,47

Los valores de ECM obtenidos para el Caso I y el Caso IV pueden considerarse como límites inferiores y superiores, respectivamente del Caso II. Las observaciones del Caso I no están contaminadas con errores sistemáticos mientras que las mediciones del

Caso IV sí. Debido a que la metodología propuesta no se ha utilizado para el último caso, la presencia de ESPT deteriora la solución del problema de RDR y el ECM aumenta. Además, los valores de ECM para el Caso II muestran que las estimaciones de las variables son significativamente mejores que las alcanzadas en el Caso IV. Esta misma tendencia se observa cuando se compara el Caso III con el Caso V. La comparación del Caso IV y Caso V muestran que a este último le corresponde menor ECM, esto se debe a que los ESPT simulados son de menor magnitud que los correspondientes al Caso IV.

La comparación de la Tabla 5.4 y 5.7 muestra que para igual *N* se alcanza un menor $\%DT_{ESPT}$ en el Caso III. Esto se debe a la menor magnitud del error considerada en el último caso. Asimismo, las Tablas 5.5 y 5.8 coinciden en que a mayor *N* los índices individuales mejoran, con excepción de $\%DT_{ESE}$. Finalmente, las Tablas 5.6 y 5.9 muestran que para el Caso III se necesita una mayor cantidad de mediciones con ESPT para declarar a una variable como sospechosa. En los casos estudio considerados se alcanzan medidas de desempeño aceptables para $N>30$.

5.6.2 Red de Intercambiadores de Calor (HEN)

El funcionamiento del HEN está representado por 17 ecuaciones de balance de masa y energía, que comprenden 16 variables medidas y 14 no medidas (Fig 4.7). Los desvíos estándares de los caudales y las temperaturas son del 2% de sus valores reales y 0,75K, respectivamente. En este caso de estudio se simularon 136 ESPT.

Las medidas de desempeño global para los Casos II y III se presentan en los cuadros 5.11 y 5.14. Los índices individuales se muestran en los cuadros 5.12 y 5.15, el $\hat{\theta}_d$ para los diferentes tipos de ESPT se incluyen en los cuadros 5.13 y 5.16 y el cuadro 5.17 presenta el ECM para todos los casos estudios.

Tabla 5.11. Índices de Desempeño Global vs *N* - Caso II (HEN)

N	% DT_{ESPT}	% FA_{ESPT}	%DT	ECM
26	83,09	24,67	90,19	109,746
30	88,97	22,93	90,97	55,094
32	99,26	0,00	97,72	0,222
40	99,26	0,74	97,23	0,266

Al igual que en el Caso II del proceso anterior, los índices de desempeño se ven favorecidos con *N* y el ECM tiene relación inversa al $\%DT_{ESPT}$. Se observa que el ECM alcanza su mínimo en *N*=32 cuando se obtiene la mayor detección de ESPT y no se presentan falsas alarmas. Esta ventana tiene igual $\%DT_{ESPT}$ que *N*=40, por lo que se analiza los índices de desempeño de ambas. La comparación entre *N*=32 y *N*=40, muestra que aun cuando ambas ventanas alcanzan igual $\%DT_{ESPT}$, el aumento en el $\%FA_{ESPT}$ provoca la disminución del $\%DT_{ESE}$ y %DT.

Tabla 5.12. Índices de Desempeño Individual vs *N* - Caso II (HEN)

N	% DT_{ESE}	%BD	%BCC	%BCE	%DD	%DCC	%DCE
26	90,48	84,29	80,00	4,29	81,82	80,30	1,52
30	91,05	87,14	81,43	5,71	90,91	89,39	1,52
32	97,66	98,57	92,86	5,71	100	100	0
40	97,15	98,57	92,86	5,71	100	100	0

Respecto de los $\hat{\theta}_d$ se observa que para $N \geq 32$, todos los BIs son detectados 4 mediciones después de que se inicie el ESPT, mientras que las DEs necesitan entre 6-8 observaciones para declarar a la variable como sospechosa. Esto se debe a que las primeras mediciones con DEs no generan estadísticos que superen el crítico y por lo tanto

los primeras observaciones con errores sistemáticos no son detectados hasta que el error está más desarrollado.

Tabla 5.13. Tiempos de Detección vs *N* - Caso II (HEN)

N	θ_d ,BCC			θ_d ,BCE			θ_d ,DCC			θ_d ,DCE		
	$\hat{\theta}_d$	*Nv*	$\bar{\theta}_{d,Nv}$	$\hat{\theta}_d$	*Nv*	$\bar{\theta}_{d,Nv}$	$\hat{\theta}_d$	*Nv*	$\bar{\theta}_{d,Nv}$	$\hat{\theta}_d$	*Nv*	$\bar{\theta}_{d,Nv}$
26	4	1	46	4			7	12	8,75	74	1	74
30	4	1	31	4			6	24	7,75	54	1	54
32	4			4			6	31	8			
40	4			4			6	27	7,56			

El Caso III alcanza el mejor desempeño para *N*=50. Contrariamente para *N*<32, la detección de ESPT es menor al 75%, por lo que no se consideran ventanas de menor longitud.

Tabla 5.14. Índices de Desempeño Global vs *N* - Caso III (HEN)

N	% DT_{ESPT}	% FA_{ESPT}	%DT	ECM
32	80,88	37,14	89,32	44,602
40	93,38	16,45	96,22	0,511
50	96,32	3,68	97,06	0,267
60	96,32	2,24	96,74	0,270

Al igual que en el Caso III del proceso lineal, el ECM disminuye a medida que *N* y $\%DT_{ESPT}$ aumentan, sin embargo, cuando *N*=60 esta tendencia cambia. Las ventanas con *N*=50 y *N*=60 tienen igual $\%DT_{ESPT}$, no obstante, la primera alcanza el menor ECM. Los índices individuales de BIs y DEs, presentan una única diferencia que se encuentra en el

%BCE. En *N*=50 se tiene un mayor %BCE, por lo que más BIs son tratados como DEs. Esta situación es similar a la que se presenta en el Caso III del proceso anterior.

Tabla 5.15. Índices de Desempeño Individual vs *N* - Caso III (HEN)

N	% DT_{ESE}	%BD	%BCC	%BCE	%DD	%DCC	%DCE
32	89,67	68,57	62,86	5,71	93,94	92,42	1,52
40	96,34	87,14	81,43	5,71	100	100	0
50	97,09	92,86	85,71	7,14	100	100	0
60	96,76	92,86	87,14	5,71	100	100	0

Table 5.16. Tiempos de Detección vs *N* - Caso III (HEN)

N	$\theta_{d,BCC}$			$\theta_{d,BCE}$			$\theta_{d,DCC}$			$\theta_{d,DCE}$		
	$\hat{\theta}_d$	*Nv*	$\bar{\theta}_{d,Nv}$	$\hat{\theta}_d$	*Nv*	$\bar{\theta}_{d,Nv}$	$\hat{\theta}_d$	*Nv*	$\bar{\theta}_{d,Nv}$	$\hat{\theta}_d$	*Nv*	$\bar{\theta}_{d,Nv}$
32	4	3	6	4			8	27	14,48	35	1	35
40	4	5	6,20	4			7	32	9,34			
50	4	7	7	4			7	26	8,85			
60	4	8	7,38	4			7	21	8,76			

La Tabla 5.17 muestra la reducción en los valores de ECM obtenidos aplicando la metodología propuesta. Se observa que la estrategia consigue reducir el ECM del proceso no lineal utilizado.

Tabla 5.17. ECM vs *N* (HEN)

N	Caso I	Caso II	Caso III	Caso IV	Caso V
26	0,034	109,746	--	2.042,568	--
30	0,030	55,094	--	2.085,60	--
32	0,028	0,222	44,602	1.928,193	811,457
40	0,022	0,266	0,511	1.739,20	862,34
50	0,018	0,309	0,267	1.507,10	1.405,92

La comparación realizada en ambos procesos muestra que:

- Caso II: obtienen medidas de desempeño satisfactorias con $N \geq 32$.
- Caso III, se obtienen medidas de desempeño satisfactorias para $N=50$.
- Los casos de estudios analizados detectan el 80% de las mediciones con BIs cuando $N \geq 40$,
- Para $N \geq 40$ todos las DEs logran ser correctamente clasificadas
- Para $N \geq 40$ el θ_d de los errores simulados en los procesos no superan nunca $N/2$

La metodología propuesta se ha aplicado a otros casos de estudios. Por razones de espacio, sólo se incluyen algunos comentarios sobre los resultados obtenidos en esos trabajos a continuación:

1. El desempeño de la estrategia para abordar procesos con corrientes paralelas se ha analizado en procesos cuya operación es representada por sistemas de ecuaciones lineales. En estos se obtienen altos porcentajes de detección y clasificación correcta cuando los ESPT se generan con la misma probabilidad para todas las observaciones. El mismo comportamiento se observa si los ESPT se colocan simultáneamente en las corrientes paralelas.

2. Se evalúa el porcentaje de detección, clasificación correcta, % DT_{ESPT} y ECM cuando sólo se simula la presencia de BIs en una sola variable medida y luego se hace lo mismo con las DEs. Las magnitudes del BI y DE varían en el rango [0 - 10] y [0 – 2]. Los resultados muestran que las medidas de desempeño para cada variable medida se favorecen con el índice de redundancia de cada variable.

5.7. Conclusiones

En este capítulo se presenta una nueva metodología para la detección y clasificación de ESPT. Su adecuada utilización en conjunto con la RDR mejora significativamente la exactitud de las estimaciones de las variables. El desempeño del algoritmo propuesto se analiza utilizando dos procesos extraídos de la literatura de RDC. A diferencia de trabajos previos, el tipo de error sistemático y el tiempo en que se presentan se simulan al azar. Asimismo, se consideran magnitudes de error más bajas que las hasta ahora utilizadas.

El análisis de las medidas de desempeño globales indica que el ECM disminuye con el incremento de N y del $\%DT_{ESPT}$, esto se debe a que se toman medidas correctivas para reducir el efecto perjudicial de los errores sistemáticos en la RDR. No obstante, existe un N a partir del cual las falsas alarmas o la presencia de múltiples errores sistemáticos afectan la RDR y y provocan un leve aumento del ECM.

Los porcentajes individuales de detección y clasificación son útiles para mostrar el desempeño de los procedimientos TRM y RLR, respectivamente. Se observa que la detección y clasificación correcta de BI y DE mejoran con N. Esto se debe a que una mayor cantidad de observaciones permiten mejorar el desempeño de los test. Por otro lado, $\%DT_{ESE}$ y %DT disminuyen con el $\%FA_{ESPT}$ y con el $\%DT_{ESPT}$. Esto se debe a que durante el transcurso de un ESPT no se efectúa el TRM, con lo cual una menor cantidad de ESE es detectada. Sin embargo, el efecto de los ESE logra ser reducido por la RDR por esto al momento de elegir una ventana resulta más relevante analizar el desempeño de los índices de ESPT.

El ECM representa la exactitud de las estimaciones de las variables. Sin embargo, todas las medidas de desempeño están vinculadas, ya que el ECM no solo varía con el $\%DT_{ESPT}$ y el $\%FA_{ESPT}$, sino también con ESPT clasificados erróneamente. Por esta

razón, no sólo la detección de ESPT es importante, sino también la clasificación que permite realizar acciones correctivas adecuadas sobre las mediciones.

Un 80% de los BI simulados logran ser detectados, con lo cual un 80% de las veces que este error se presente, las observaciones podrán ser corregidas y estar disponibles hasta que el sensor sea reparado.

Finalmente, para los procesos y casos analizados, se obtienen altos porcentajes de detección y clasificación para valores de N en el intervalo [32 - 50]. En base a estos resultados, se puede concluir que $N = 40$ es una longitud de ventana apropiada para aplicar acciones correctivas rápidas y lograr elevadas medidas de desempeño.

5.8 Notación

$\mathbf{a}$	Vector de ajuste de las mediciones
B	Magnitud del sesgo
df	Grado de libertad de la distribución de Student
$\mathbf{f}$	Sistema de restricciones de igualdad
F	Función de distribución acumulada ideal
G	Función de distribución acumulada no ideal
$\mathbf{h}$	Sistema de restricciones de desigualdad
I	Número de variables medidas
K	Magnitud del ESE
m_{drift}	Magnitud de la deriva
N	Número de réplicas de la variable medida
Ns	Número de simulaciones
Nv	Número de veces que $\theta_d > \tilde{\theta}_d$
p_G	Probabilidad global de errores sistemáticos
$\mathbf{Q}$	Matriz de covarianza del ajuste
s	Parámetro del algoritmo
T_β	Estadístico de test de la pendiente
$\mathbf{u}$	Vector de variables no medidas
$\hat{\mathbf{u}}$	Vector estimado de las variables no medidas
$\mathbf{u}^U$	Límite superior de las variables no medidas
$\mathbf{u}^L$	Límite inferior las variables no medidas
U	Número de variables no medidas
W	Función de peso del M-estimador

$\mathbf{y}$	Vector de mediciones
$\tilde{y}$	Mediana robusta de las observaciones
$\mathbf{Y_{ob}}$	Matriz de observaciones
y_s	Conjunto de mediciones dependientes para la regresión lineal
$\mathbf{x}$	Vector de variables medidas
$\hat{\mathbf{x}}$	Vector reconciliado de las variables medidas
$\mathbf{x^U}$	Límite superior de las variables medidas
$\mathbf{x^L}$	Límite inferior de las variables medidas
x_s	Conjunto de mediciones independientes de la regresión lineal
α	Nivel de significancia del test
$\hat{\beta}$	Parámetros de la regresión
ρ	M-estimador
ψ	Función de influencia del M-estimador
ξ	Fracción de datos que no corresponden a la distribución ideal
η	Parámetro del algoritmo
$\boldsymbol{\varepsilon}$	Vector de errores aleatorios
ω	Parámetro del algoritmo
$\mathcal{V}$	Varianza de la M-estimación
σ_r	Desvio estándar del residuo de la regresión
σ_y	Desvio estándar de la medición
τ	Estadístico
τ_c	Estadístico crítico
$\hat{\mathcal{J}}$	M-estimación de $\mathcal{J}$

θ_d	Tiempo de detección
$\tilde{\theta}_d$	Mediana de los tiempos de detección
$\bar{\theta}_{d,Nv}$	Media de los valores de θ_d para los cuales se verifica que $\theta_d > \tilde{\theta}_d$

5.9 Acrónimos

%BCC	Porcentaje de sesgos Correctamente Clasificados
%BD	Porcentaje de sesgos Detectados
%BEC	Porcentaje de Sesgos Erróneamente Clasificados
BI	Sesgo
BW	Función Biweight
CM	Cuadrados Mínimos
CO	Función Correntropía
%DCC	Porcentaje de Derivas Correctamente Clasificadas
%DD	Porcentaje de Derivas Detectadas
DE	Deriva
%DEC	Porcentaje de Derivas Erróneamente Clasificadas
%DT	Porcentaje de Detección Total de Errores Sistemáticos
$\%DT_{ESE}$	Porcentaje de Detección Total de ESE
$\%DT_{ESPT}$	Porcentaje de Detección Total de ESPT
ECM	Error Cuadrático Medio
Ef	Eficiencia Asintótica
ESE	Error Sistemático Esporádico
ESPT	Error Sistemático que Persiste en el Tiempo
ERTP	Estimador Redescendiente en Tres Partes

$\%FA_{ESPT}$	Porcentaje de Falsas Alarmas de ESPT
HU	Función de Huber
HEN	Red de Intercambio de Calor
MSi	Método Simple
MAD	Mediana de los desvíos absolutos alrededor de la mediana
PQ	Punto de Quiebre
RDC	Reconciliación de Datos Clásica
RDR	Reconciliación de Datos Robusta
RLR	Regresión Lineal Robusta
SMN	Red de Ingreso de Vopr
TRM	Test Robusto de las Mediciones
WE	Función de Welsch

Capítulo 6

Aplicación a la Producción de Biodiésel

6 Aplicación a la Producción de Biodiésel

6.1 Introducción

El desempeño del procedimiento propuesto en el Capítulo 5 se estudia para un sistema de mayor escala. Con este fin se selecciona un proceso actual relevante para nuestro país como es la producción de Biodiésel, la cual se proyecta en 3,05 millones de m^3 para el presente año (Kenneth J., 2017). Éste se obtiene principalmente de aceite de soja, no obstante, existen programas, por ejemplo, el incentivado por el Ministerio de Agroindustria de la Provincia de Buenos Aires, que promueven la reutilización de aceites usados para la producción de dicho combustible. Por esto se modela una planta que produce 10 mil m^3 por año, sobre la cual se aplica la metodología desarrollada para el tratamiento general de los errores de las mediciones con vistas a su optimización en línea.

6.2 Descripción General

6.2.1 Producción en la Argentina

Desde principios de 2007, la Argentina ha implementado la Ley 26.093 de Biocombustibles, ésta establece que a partir de 2010 la nafta se mezcle con bioetanol y el diésel con biodiésel. Originalmente los cortes propuestos fueron del 5% (Artículo 7. Ley 26.093), después de varias modificaciones, el diésel se encuentra en el 10% y la nafta en 12%. Para este año se prevé que la producción de biodiésel para el mercado interno aumente a 1,35 millones de m^3, la cifra más alta hasta la fecha.

El negocio del biodiésel se inició cuando las grandes plantas locales de producción de aceite vegetal vieron la oportunidad de agregar valor a sus productos y comenzaron a exportar biodiésel a la Unión Europea. Estas exportaciones respaldaron la mayor parte

del crecimiento de la industria hasta alcanzar su máximo en 2011. Actualmente, el principal mercado es Estados Unidos, seguido de Perú. Prácticamente todo el biodiésel producido en la Argentina está hecho de aceite de soja, mientras que hay un volumen poco significativo de biodiésel producido a partir de aceite de cocina usado.

El biodiésel es producido por grandes procesadores que utilizan plantas de trituración de oleaginosas totalmente integradas. Mientras que las plantas pequeñas y medianas, compran aceite vegetal a las más grandes. La mayoría de las plantas pequeñas operan con alta capacidad para cumplir con el requerimiento del mercado local, mientras que las plantas grandes operan a media capacidad y se centran casi exclusivamente en el mercado de exportación.

Hay 53 plantas de biodiésel inscriptas en el Ministerio de Energía y Minería, las que tienen una capacidad de producción de hasta 0,7 millones de m^3 por año. Las diez compañías más grandes representan más del 70% de la capacidad del país. Éstas pertenecen a empresas con inversores internacionales y locales que cuentan con grandes instalaciones de trituración de semillas oleaginosas en el país. El saldo se distribuye entre empresas más pequeñas, con plantas con capacidades de producción que oscilan entre 12 y 110 miles de m^3 por año. Basado en este análisis de la Producción se desarrolla el modelado simplificado de una planta de biodiésel que produce 10 mil m^3/año.

6.2.2 Reacción química básica

El biodiésel, también conocido como FAME por su sigla en inglés: *Fatty Acids Methyl Esters,* es un combustible renovable de características similares al gasoil, producido a partir de aceites vegetales o grasas animales. Los mismos están constituidos principalmente por triglicéridos, que pueden reaccionar con alcohol en presencia de un

catalizador en una reacción conocida como transesterificación, dando como productos ésteres monoalquílicos de ácidos grasos de cadena larga. Esta reacción está formada por tres etapas que dan como resultado tres moles de alquil éster por mol de triglicérido (Fig.6.1).

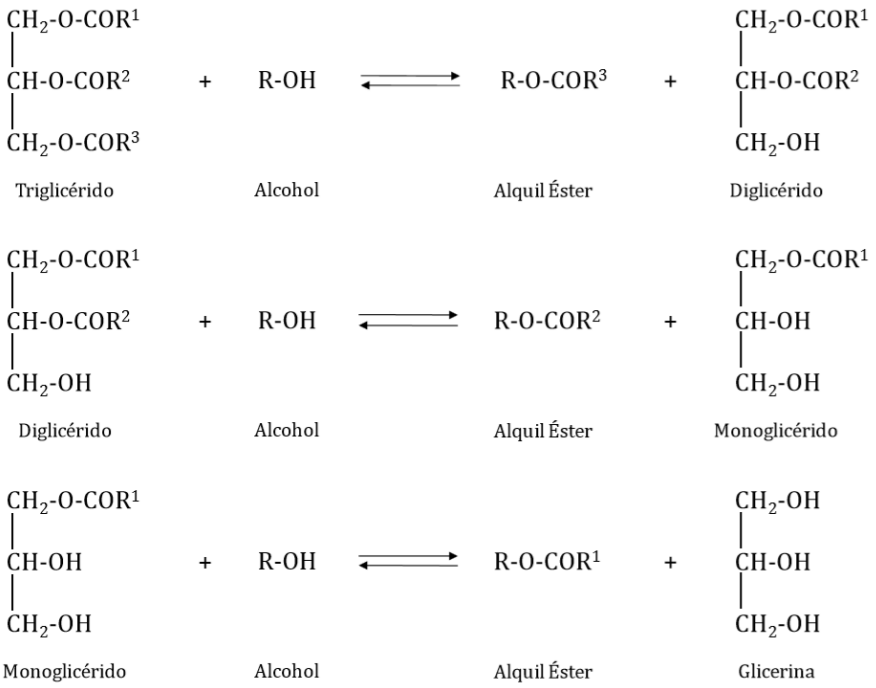


Figura 6.1: Transesterificación de triglicéridos para la síntesis de alquil ésteres y subproductos (glicerina, monoglicérido y diglicérido)

donde R^1, R^2 y R^3 representan las cadenas de ácidos grasos, que pueden ser distintas o iguales. El alcohol utilizado en esta reacción, por lo general, es el metanol debido a su bajo costo. A continuación, se presenta la forma simplificada de esta reacción, cuando las cadenas de ácidos grasos son iguales (Fig.6.2).

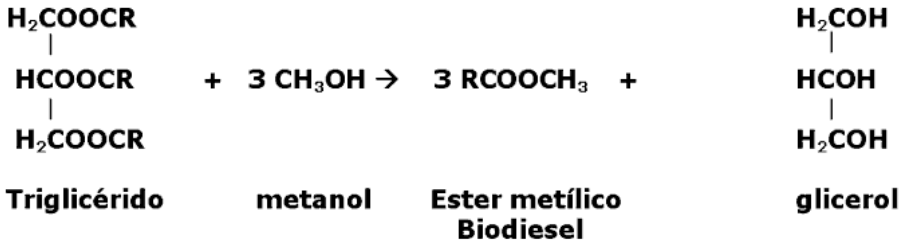


Figura 6.2: Reacción de obtención de biodiésel simplificada

Por lo general esta reacción se realiza con grandes excesos de metanol a fin de propiciar que el equilibrio se encuentra desplazado hacia la derecha.

A escala industrial, la reacción de transesterificación se realiza en presencia de un catalizador químico homogéneo básico, a temperaturas entre 37ºC y 70ºC y presión atmosférica. Estos catalizadores son ampliamente utilizados por su alto rendimiento a bajas temperaturas y bajo costo (Hegel y co, 2008). Sin embargo, dicho proceso presenta la desventaja de ser altamente sensible a la presencia de impurezas en los reactivos, produciendo compuestos no deseados como jabones, mediante la reacción de saponificación (Cotabarren, 2017). En consecuencia, se busca que la materia prima tenga un bajo contenido de agua y ácidos grasos libres (Vyas y co., 2010), para que la reacción transcurra en forma eficiente y minimizar así las pérdidas en el proceso.

En procesos que utilizan materias primas de menor calidad, con alto contenido de ácidos grasos libres, como es el caso de aceites usados, se utilizan catalizadores homogéneos ácidos (Canacki y Gerpen, 2001; Lotero y co., 2005). El uso de estos reactivos sin refinar tiene como desventaja la necesidad de contar con un mayor número de etapas de separación y neutralización de los catalizadores, incrementando los costos de procesamiento (Kulcarni y Dalai , 2006). No obstante, la utilización de residuos de

aceite con el fin de producir biodiésel es promovida por las políticas de gobierno y grupos ambientalistas.

6.2.3 Sistema Ácido-Catalítico

El estudio de la obtención de biodiésel por catálisis ácida ha sido muy limitado, esta vía ofrece beneficios gracias a su independencia del contenido de ácidos grasos libres y la consecuente ausencia de una etapa de pretratamiento. Estas ventajas favorecen el uso del proceso catalizado por ácido cuando se utilizan aceites de cocina usado u otros aceites de baja calidad como materia prima. Sin embargo, su velocidad de reacción es relativamente lenta en comparación con la catálisis básica por lo que en la industria el proceso básico es el elegido. A continuación, se presentan algunos aportes realizados para el sistema catalizado con ácido:

Freedman y co. (1984) investigaron la transesterificación de aceite de soja con metanol usando 1% en peso de ácido sulfúrico concentrado. Afirmaron que las reacciones directa e inversa seguían una cinética de pseudo primer orden y de segundo orden, respectivamente. Asimismo, estos autores también descubrieron que a 65 ° C y una relación molar de 30: 1 de metanol a aceite, se obtenía 90% de conversión de aceite en ésteres metílicos al cabo de 69 h.

Ripmester (1998) y Mcbride (1999) analizaron la cinética de la reacción ácida a escala laboratorio cuando se empleó como materia prima aceite de cocina. Además, llevaron a cabo reacciones de transesterificación a escala piloto, en un reactor de acero inoxidable de 15 L equipado con una camisa calefactora y agitador, el cual operó a 400 rpm. La reacción se desarrolló a 70 ° C y una presión de 170-180 kPa con exceso de metanol (relación molar mínima de 50:1 de metanol a aceite) y en presencia de ácido

sulfúrico, con concentraciones en el rango 1,5-3,5% en moles. Las proporciones de metanol fueron elevadas con el fin de promover altas conversiones de aceite a éster. Bajo estas condiciones, se alcanzó una conversión del 97% de aceite a FAME en 240 min. Estos autores propusieron un modelo empírico de primer orden y calcularon una constante de velocidad, la cual fue utilizada para dimensionar los equipos de los diagramas de flujos presentados en Zhang y co. (2003).

La reacción ácida también fue estudiada por Canakci y Gerpen (2001), quienes analizaron los efectos de la relación molar de alcohol a aceite de soja, la temperatura de reacción, la cantidad de catalizador y el tiempo de reacción en la conversión de éster mediante transesterificación. Los efectos mencionados se estudiaron por separado. Estos autores concluyeron que la conversión de éster incrementa con la relación molar de alcohol a aceite, así como también con el aumento de la temperatura de reacción, la concentración de ácido sulfúrico y el tiempo de reacción. Sin embargo, la posible interacción de estas variables no fue analizada y tampoco se recomendó una condición óptima para la reacción.

Al-Widyan y Shyoukh (2002) obtuvieron biodiésel a partir de aceite de palma usado. La reacción de transesterificación se probó con etanol y 2 catalizadores ácidos, HCl y H_2SO_4. Este último resultó ser el más efectivo cuando su concentración se encuentra en el rango 1,5:2,5 molar. Además, lograron reducir notablemente el tiempo de reacción con exceso de alcohol.

Zheng y co. (2006) analizaron los efectos de mezclado, composición de la alimentación y temperatura, siendo estos dos últimos los factores más significativos en el rendimiento de la reacción. El estudio permitió concluir que, en exceso de metanol, la reacción puede ser modelada como una reacción de pseudo primer orden para

temperaturas en el rango de 70-80 °C y con relaciones de metanol aceite 74:1 - 250:1. La relación aceite: metanol: ácido de 1:74:1.9 a 70°C permitió alcanzar conversiones del 98.9 % siendo una de las más prometedoras debido a la menor cantidad de metanol involucrada.

El biodiésel obtenido a partir de aceite de cocina usado ha demostrado tener mejor desempeño y producir menores emisiones de hidrocaruros y CO cuando se lo usa en motores diésel (Kulkarni y Dalai, 2006). Por esto Lam y co. (2010) presentaron una revisión en la que mencionan la importancia de buscar alternativas que permitan utilizar el aceite usado como materia prima para la producción de biodiésel y detallan las ventajas y desventajas de cada catalizador utilizado para tal fin.

6.3 Modelo del Proceso

Se realizó un modelo simplificado de una planta de producción de biodiésel por catálisis ácida. Este se formuló siguiendo el trabajo de Zhang y co. (2003), para una producción anual de 10.000 m^3. Las principales unidades que este proceso incluye son: un reactor de transesterificación, un tren de columnas de separación, intercambiadores de calor, bombas y separadores. Se asume que el reactor opera de forma continua y es un reactor tanque agitado que trabaja con medio volumen vacío. Se utilizan múltiples etapas de destilación para la recuperación del metanol, purificación del FAME y separación de la glicerina. Cabe destacar que el FAME se separa de la mezcla glicerol-metanol por medio de una extracción líquido-líquido. En la Fig. 6.3 se presenta el diagrama de la planta. A continuación, se describen las condiciones de operación de las principales unidades.

6.3.1 Reacción de Transesterificación

Las corrientes de alimentación al reactor se ajustan a una relación molar de 50:1.3:1 de metanol: ácido sulfúrico y aceite usado. Las condiciones dentro del reactor se fijan en 80ºC de temperatura y una presión de 400 kPa.

Tres corrientes que contienen metanol fresco (corriente F_1, 216 kg / h), metanol reciclado (corriente F_8, 1594 kg / h) y ácido sulfúrico (corriente F_3, 150 kg / h) se mezclan y luego son alimentadas al reactor de transesterificación (**E-9**) mediante una bomba. El aceite de cocina usado (corriente F_{10}, 1030 kg / h) entra en **E-9** a una temperatura de 60 °C, la cual se alcanza con un intercambiador de calor (**E-8**) por el cual circula la corriente de salida del reactor F_9. Se asume que el 97% del aceite se convierte a FAME, después de 4 h de reacción. Además, se considera que la reacción que se presenta en el reactor es de pseudo primer orden y que el aceite sólo contiene cadenas de ácidos grasos correspondientes al ácido oleico ($-C_{18}H_{34}O_2$).

6.3.2 Recuperación de Metanol

Debido al gran exceso de metanol en la corriente F_9, la recuperación de este reactivo es el primer paso después de la reacción. De esta forma se disminuye la carga en las unidades aguas abajo. En la columna de destilación de metanol **E-10**, se alcanza una tasa de recuperación de metanol del 94%. La corriente de destilación F_{15} se recircula a **E-9**, mientras que la corriente inferior F_{16} se envía a la unidad de neutralización y eliminación del ácido **E-12**.

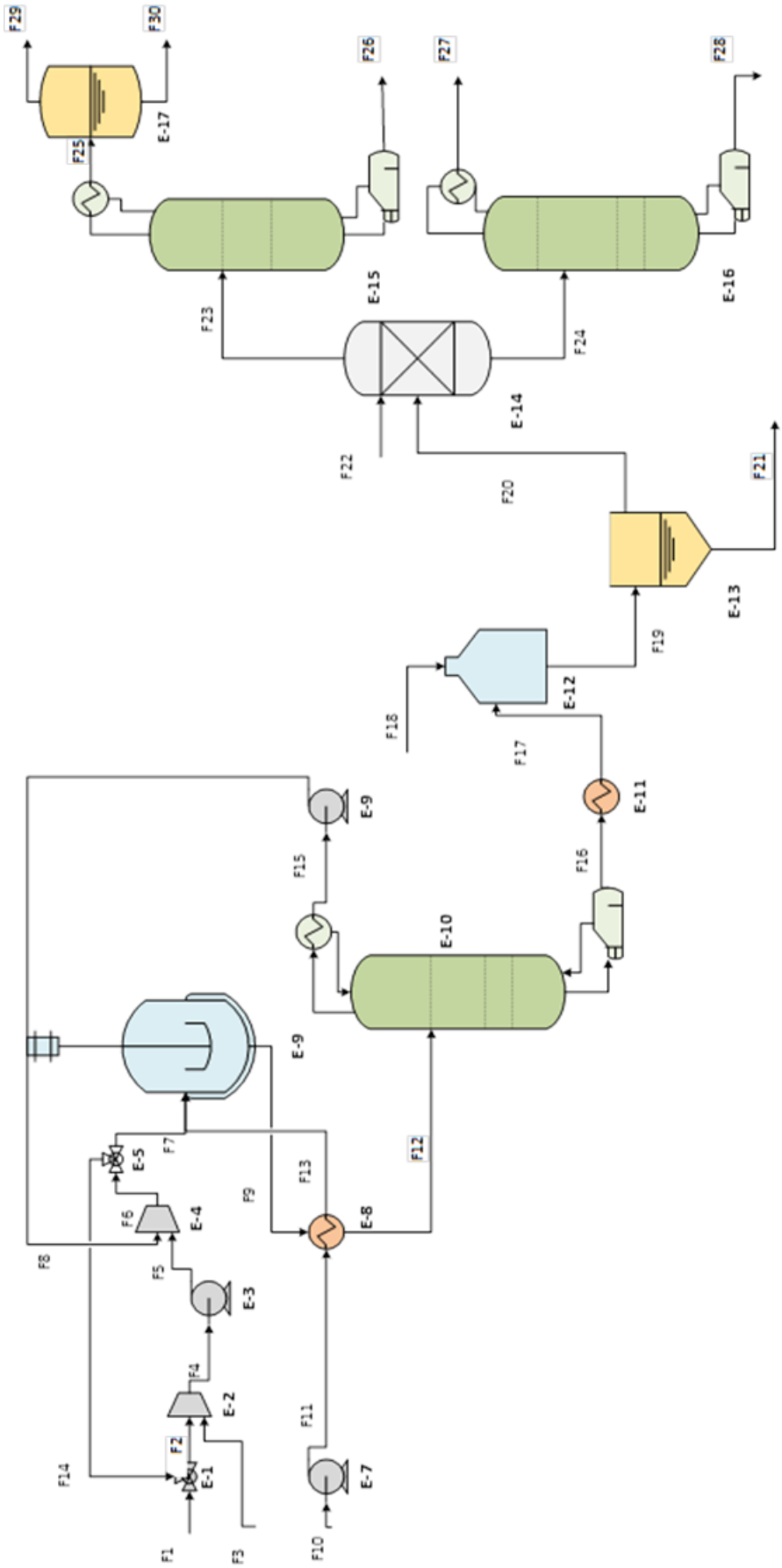


Figura 6.3. Diagrama del Proceso de Producción de Biodiésel

6.3.3 Eliminación del ácido

En el reactor **E-12**, el ácido sulfúrico se elimina completamente en una reacción de neutralización añadiendo óxido de calcio (CaO) para producir $CaSO_4$ y H_2O. El óxido de calcio se utiliza como base debido a su bajo costo en relación con otras sustancias alcalinas. La reacción de neutralización que se produce es:

$$CaO + H_2SO_4 \rightarrow CaSO_4 + H_2O$$

El agua producida es absorbida por el $CaSO_4$ producido para formar $CaSO_4.2H_2O$. La eliminación del $CaSO_4.2H_2O$ se realiza en un separador por gravedad, **E-13**. La corriente sobrenadante F_{20} (1247 kg / h) contiene 79% de FAME, 9% de glicerol, 8% de metanol y 2% de aceite no convertido de porcentaje en masa. Dicha corriente se envía a la columna **E-14,** donde por medio de una corriente de agua F_{22} se produce la separación entre el FAME y el glicerol.

6.3.4 Ecuaciones de balance de proceso

El modelo está formado por 127 variables que corresponden a flujos, temperaturas y composiciones. Éstas pueden clasificarse como medidas en línea, medidas con retardo y no medidas. Se formulan 115 ecuaciones que incluyen 17 balances totales de masa, 11 balances de energía, 63 balances por componente, 3 relaciones de equilibrio líquido-líquido y 21 ecuaciones de normalización. A continuación, se presentan las expresiones generales de los balances realizados en los equipos.

- Balances de Flujos para equipos sin reacción

Se realizan balances de masa en los 17 equipos que forman parte del proceso. Éstos se encuentran representados por la siguiente ecuación:

$$\sum_{i=1}^{n} F^{E}{}_{i} - \sum_{i=1}^{n} F^{S}{}_{j} = 0 \quad i \neq j; i = 1:30;\ j = 1:30 \tag{6.1}$$

donde *i* y *j* representan los flujos involucrados en el proceso.

- Balances por componentes para equipos sin reacción

Se formulan 50 balances sin reacción química que son descriptos por la ecuación:

$$\sum_{i=1}^{n} F^{E}{}_{i} x_{i,c} - \sum_{i=1}^{n} F^{S}{}_{j} x_{j,c} = 0 \quad i \neq j \tag{6.2}$$

siendo *c*:

M	Metanol	HS	Ácido Sulfúrico
T	Triacilglicérido	A	Agua
F	FAME	C	Óxido de Calcio
G	Glicerol	CS	Sulfato de Calcio

- Balances por componente para el reactor **E-9**

$$\begin{gathered} \sum_{i=F_7}^{F_{13}} F^{E}{}_{i} x_{i,c} - F^{S}{}_{9} x_{9,c} + \upsilon r V = 0 \\ i = F_7, F_{13};\ c = [M, T, F, G, HS] \end{gathered} \tag{6.3}$$

siendo υ es el coeficiente estequiométrico de la reacción y la cinética está descripta por:

$$rV = k C_{T,9} V = \ddot{\mathrm{K}} F_9 x_{T,9} \tag{6.4}$$

donde $\ddot{\mathrm{K}}$ representa el agrupamiento de las constantes que permiten obtener la expresión final.

- Balances con reacción en **E-12**

Se asume que todo el CaO alimentado reacciona con el ácido sulfúrico presente en la corriente de entrada. Por lo tanto, los balances de las especies reaccionantes se representan como:

$$\frac{F^{E}{}_{17}x_{17,c}}{PM(c)}+\upsilon\frac{x_{CaO}F_{18}}{PM(CaO)}-\frac{F^{s}{}_{19}x_{19,c}}{PM(c)}=0$$
$$c=[HS,A,CS] \tag{6.5}$$

- Balances de Energía

Se realizan 11 balances de energía en equipos sin reacción

$$\sum_{i=1}^{n}F^{E}{}_{i}H_{i}(P,T)-\sum_{i=1}^{n}F^{S}{}_{j}H_{j}(P,T)+Q_{e}^{s}=0 \quad i\neq j;\, i=1{:}30;\; j=1{:}30 \tag{6.6}$$

- Balance de Energía para el reactor **E-1**

$$F_{7}\tilde{c}_{p,7}\left(T_{7}\right)\left[T_{7}-T_{r}\right]+F_{13}\tilde{c}_{p,T}(T_{13})\left[T_{13}-T_{r}\right]-F_{9}\tilde{c}_{p,9}(T_{9})\left[T_{9}-T_{r}\right]+\Delta H_{r}(T)=0 \tag{6.7}$$

donde ΔH_r representa el calor de reacción.

6.4 Modelo de las mediciones

Las observaciones con errores aleatorios pueden describirse con el siguiente modelo:

$$y_{ij}=x_{i}+\varepsilon_{ij}, \tag{6.8}$$

donde x_i representa el valor verdadero de la variable y ε_i representa el error aleatorio, el cual se distribuye según una normal $\mathcal{N}(O,\sigma_y)$ siendo σ_y la varianza de ε_i. Asimismo, los errores sistemáticos esporádicos (ESE) y errores que persisten en el tiempo (ESPT) están descriptos por los siguientes modelos:

$$y_{ij} = x_i + \varepsilon_{ij} + K_{ij}\sigma_{y,i}\,, \tag{6.9}$$

$$y_{ij} = x_i + \varepsilon_{ij} + \mathrm{B}_{ij}\sigma_{y,i}\,, \tag{6.10}$$

$$y_{ij} = x_i + \varepsilon_{ij} + m_{ij,drift}\ t\ \sigma_{y,i}\,. \tag{6.11}$$

donde la magnitud de los ESE se representa con K_{ij}, la del sesgo (BI) con B_{ij} y de la deriva (DE) se representa como m_{drift}. La persistencia de errores sistemáticos en el tiempo ocasiona el deterioro de las estimaciones y puede conducir al proceso a operar en condiciones inseguras, poco eficientes o no obtener el producto final deseado

El estudio realizado en el Capítulo 5 sobre procesos de pocas variables permitió concluir que los ESPT detectados a tiempo pueden ser tratados por medio de una estrategia de detección y clasificación basada en Estadística Robusta. Esta estrategia permite tomar acciones correctivas sobre las mediciones según el tipo de error, de manera que la Reconciliación Datos Robusta (RDR) proporcione estimaciones insesgadas. En la siguiente figura se muestra un esquema simplificado de la metodología desarrollada en el Capítulo 5.

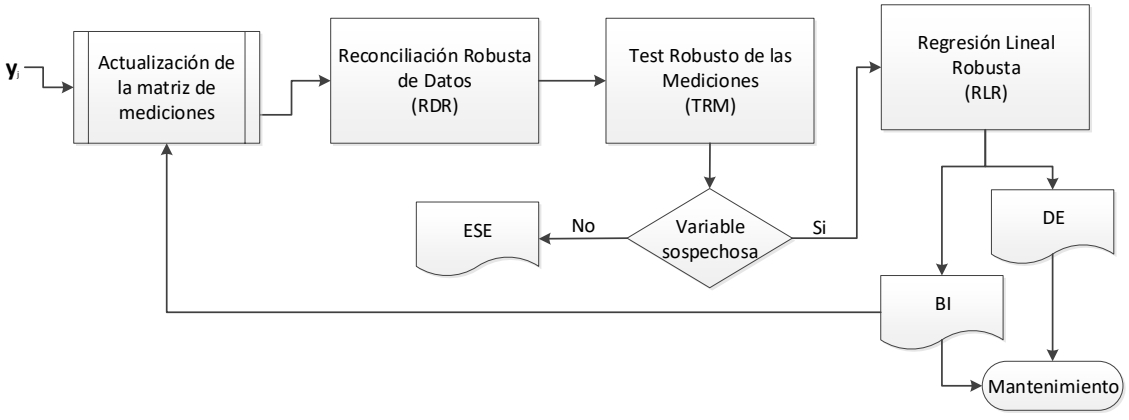


Figura 6.4 Esquema simplificado de la metodología desarrollada en el Capítulo 5

Las plantas de producción de biodiésel cuentan con instrumentación que les permite hacer el monitoreo en línea de sus variables. Esto se debe a que se busca obtener un producto final de una calidad definida y hacer un uso óptimo de la materia prima. Por esto se procede a probar la estrategia desarrollada en un modelo que describe la producción del biodiésel.

6.5 Análisis del desempeño

Se generan mediciones atípicas en el conjunto de variables medidas en línea solamente. Dada una probabilidad de ocurrencia de errores sistemáticos, el 95% de las veces se generan ESE y el 5% restante se simulan ESPT. Estos últimos pueden ser BI o DE en igual proporción y se mantienen durante 100 instante de tiempo consecutivos. Asimismo, se considera que un sensor reparado no vuelve a presentar ESPT durante 400 mediciones.

Las medidas de desempeño seleccionadas para analizar del comportamiento de la metodología están vinculadas principalmente a la detección de ESPT. Por esto se extraen del Capítulo 5 los siguientes índices:

- Porcentaje de ESPT detectados ($\%DT_{ESPT}$):

$$\%\mathrm{DT}_{ESPT} = \frac{\left(\#ESPT\right)_{Simulados\ y\ Detectados}}{\left(\#ESPT\right)_{Simulados}} x\,100 \tag{6.11}$$

- Falsas alarmas de ESPT ($\%FA_{ESPT}$):

$$\%\mathrm{FA}_{ESPT} = \frac{\left(\#ESPT\right)_{Detectados} - \left(\#ESPT\right)_{Simulados\ y\ CorrectamenteDetectados}}{\left(\#ESPT\right)_{Detectados}} x\,100 \tag{6.12}$$

- Error Cuadrático Medio (ECM*):

$$\text{ECM*}=\frac{1}{I\ Ns}\sum_{k=1}^{Ns}\sum_{i=1}^{I}\left(\frac{\hat{x}_i-x_i}{\sigma_{y,i}}\right)^2 \tag{6.13}$$

donde *Ns* es la cantidad de simulaciones e *I* es la cantidad de variables medidas en línea. El error cuadrático medio solo se computa en las variables medidas en líneas, es por eso que este índice se indica como ECM*.

Con respecto a los índices individuales se contabiliza la relación entre los ESE que son detectados y los simulados, estos se definen como el porcentaje de ESE Detectados (% DT_{ESE}):

$$\%\text{DT}_{ESE}=\frac{(\#\,ESE)_{\,SimuladosyCorrectamenteClasificados}}{(\#\,ESE)_{\,Simulados}}\,x\,100 \tag{6.14}$$

A diferencia de los ESE, los ESPT no se detectan y clasifican simultáneamente, la detección se realiza con el TRM y la clasificación con la RLR, es por esto que se utilizan dos índices que permiten analizar el desempeño de cada uno de estos procedimientos. La detección de sesgo (BD) o de derivas (DD) son medidas del desempeño del TRM y se calculan como:

$$\%\text{XD}=\frac{(\#\,XI)_{\,Simulados\ y\ Correctamente\ Detectados}}{(\#\,XI)_{\,Simulados}}\,x\,100 \quad X\in[B;D] \tag{6.15}$$

Las Clasificaciones Correctas de estos errores se calculan como:

$$\%\text{XCC}=\frac{(\#\,XI)_{\,Correctamente\ Clasificado}}{(\#\,XI)_{\,Simulados}}\,x\,100 \quad X\in[B;D] \tag{6.16}$$

Por último, las Clasificaciones Erróneas de BIs o DEs están dadas por:

$$\%\text{XCE}=\frac{(\#\,BI)_{\,Mal\ Clasificado}}{(\#\,BI)_{\,Simulados}}\,x\,100 \quad X\in[B;D] \tag{6.17}$$

Se realizan pruebas sobre el modelo de proceso cuando las variables se clasifican según la Tabla 6.1 para ventanas de longitud *N* = [30 40 50]. Se formulan 4 casos de

estudio en los que se analiza la adecuación de la metodología desarrollada en la tesis para una planta de biodiésel. El Caso I se realiza con mediciones que solo tienen errores aleatorios, mientras que el Caso II y III consideran la presencia de ESPT generados con probabilidades de ocurrencia 3,37% y 4,75% a los que les corresponde una probabilidad de mediciones atípicas del 20 y 28%. Asimismo, se formula un Caso IV sin metodología de detección-clasificación que considera la presencia de los ESPT del Caso III.

Tabla 6.1 Clasificación de Variables

Variables	Medidas en línea	Medidas con retardo	No medidas
F	19	-	11
T	23	-	7
x_c	16	19	32

Las magnitudes de los BIs y DEs simulados varían de forma aleatoria en los rangos $B_{ij} \in [4,5-9,5]$ y $m_{drift} \in [0,75-3,25]$, mientras que los ESE tienen una magnitud fija, K=10. En la Tabla 6.2 se presentan la cantidad de errores simulados para cada caso.

Tabla 6.2 Cantidad de errores sistemáticos de cada tipo

Casos	*K*	*BI*	*DE*
Caso I	--	--	--
Caso II	27762	445	442
Caso III	53211	580	526
Caso IV	27762	445	442

6.6 Análisis de los resultados

A continuación, se presentan las métricas de detección, clasificación y ECM para la planta de biodiésel. Los índices globales del Caso II se muestran en Tabla 6.3 y sus correspondientes índices individuales en la Tabla 6.4. Asimismo, los casos analizados bajo las condiciones del Caso III se presentan en las Tabla 6.5 y 6.6 para las métricas globales e individuales, respectivamente. Por último, en la Tabla 6.7 se presentan los ECM* de todos los casos simulados.

Tabla 6.3 Índices Globales del Caso II

N	% DT_{ESPT}	% FA_{ESPT}	ECM*
30	94,93	7,06	3,725
40	98,87	2,99	0,373
50	99,32	2,33	0,297

Tabla 6.4. Índices Individuales del Caso II

N	% DT_{ESE}	%BD	%BCC	%BCE	%DD	%DCC	%DCE
30	83,64	92,58	85,84	6,74	97,29	96,61	0,68
40	83,18	97,98	93,26	4,72	99,77	99,10	0,68
50	82,24	98,88	94,16	4,72	99,77	99,10	0,68

Se observa que al igual que los procesos abordados en el Capítulo 5, el ECM* disminuye con el incremento de $\%DT_{ESPT}$ y *N*. Por otro lado, en el rango analizado, las $\%FA_{ESPT}$ disminuyen con *N*, alcanzando índices aceptables para $N>30$. Respecto de los índices individuales se observa que:

- La detección y clasificación de BIs y DEs mejoran con *N*, lo contrario sucede con el % DT_{ESE}. Este parámetro es inferior al obtenido en los procesos simulados en el Capítulo 5, sin embargo, se sabe que los ESE no deterioran la solución de la RDR.
- En todas las ventanas analizadas se presentan ESPT clasificados de forma errónea. Para N≥40, los % BCE y %DCE se mantienen constantes, en esas ventanas el test de la RLR logra clasificar de forma acertada el 97,3% de los ESPT simulados.

Los índices de desempeño del Caso III se presentan a continuación.

Tabla 6.5. Índices Globales del Caso III

N	% DT_{ESPT}	% FA_{ESPT}	ECM*
30	85,62	17,58	11,58
40	93,67	13,95	7,51
50	96,02	10,46	2,39

Tabla 6.6. Índices Individuales del Caso III

N	%DT_{ESE}	%BD	%BCC	%BCE	%DD	%DCC	%DCE
30	73,37	83,97	72,93	11,03	87,45	84,41	3,04
40	71,70	92,59	82,41	10,17	94,87	92,21	2,66
50	71,00	95,52	86,21	9,31	96,58	94,87	1,71

En la Tabla 6.5 se observa que se mantiene la tendencia de reducción del ECM* con el incremento de %DT_{ESPT} y mayor *N*. Los índices de detección y clasificación presentan el mismo comportamiento que el correspondiente al Caso II. No obstante, son inferiores a los del Caso II. Esto se debe a la presencia de una mayor cantidad de errores sistemáticos en el Caso III, que produce el deterioro de la estimación de la RDR. Esto provoca el aumento del ECM* y la disminución de desempeño del TMR, el cual eleva las falsas alarmas de ESPT y disminuye un 3% la detección de estos errores.

En la Tabla 6.7 se presenta el ECM* de los Casos II y III, así como también el ECM* obtenido cuando las mediciones solo tienen errores aleatorios (Caso I) y cuando sólo se aplica RDR a un conjunto de mediciones con errores sistemáticos (Caso IV).

Tabla 6.7. ECM* de todos los casos considerados

N	Caso 1	Caso 2	Caso 3	Caso 4
30	0,012	3,725	11,582	53,213
40	0.006	0,373	7,511	35,693
50	0.007	0,297	2,391	32,627

Se observa que el procedimiento de detección, clasificación y corrección de las mediciones consigue disminuir el ECM* de las variables medidas en línea. Asimismo, el ECM* se deteriora con el aumenta de mediciones atípicas.

6.7 Conclusiones

La metodología desarrollada puede aplicarse a sistemas en línea de mayor escala, como es el caso de una planta de producción de biodiésel. Sin embargo, las medidas de desempeño están ligadas a la cantidad de errores sistemáticos simulados. Es por esto que se obtienen mejores índices cuando dicha cantidad disminuye.

Las ventanas de longitud $N = 40$ y $N = 50$ permiten detectar un elevado porcentaje de ESPT simulados, aunque disminuyen la capacidad de detección de ESE. No obstante, estas longitudes logran reducir el ECM con lo cual las estimaciones de la RDR no se ven afectadas. Por esto estas ventanas resultan atractivas para la implementación del método en línea.

6.8 Notación

B	Magnitud del sesgo
c	Componente
$\tilde{c}_{p,i}$	Capacidad media de la corriente i
E-X	Equipo X
F^E	Flujos de Entrada
F^S	Flujos de Salida
I	Variables medidas en línea
K	Constante de velocidad
K	Magnitud del ESE
$\ddot{K}$	Constante de la expresión de velocidad de reacción
m_{drift}	Magnitud de la deriva
N	Número de réplicas de la variable medida
Ns	Número de simulaciones
PM	Peso molecular
Q_e	Calor entregado
t	Tiempo
T_i	Temperatura de la corriente i
V	Volumen del reactor
y_{ij}	Medición de la i-ésima variable en el tiempo j-ésimo
x_c	Fracción másica del componente c
x_i	valor verdadero de la i-ésima variable
ΔH_r	Entalpía de reacción
$\boldsymbol{\varepsilon}$	Vector de errores aleatorios

υ	Coeficiente estequiométrico
ω	Parámetro del algoritmo
σ_y	Desvió estándar de la medición

6.9 Acrónimos

%BCC	Porcentaje de BI Clasificados Correctamente
%BCE	Porcentaje de BI Clasificados Erroneamente
%BD	Porcentaje de BI Detectado
BI	Sesgo
%DCC	Porcentaje de DE Clasificadas Correctamente
%DCE	Porcentaje de DE Clasificadas Erroneamente
%DD	Porcentaje de DE detectados
DE	Deriva
$\%DT_{ESE}$	Porcentaje de Detección Total de ESE
$\%DT_{ESPT}$	Porcentaje de Detección Total de ESPT
ECM	Error Cuadrático Medio
ESE	Error Sistemático Esporádico
ESPT	Error Sistemático que Persiste en el Tiempo
FAME	Fatty Acids Methyl Esters
RDR	Reconciliación de Datos Robusta
RLR	Regresión Lineal Robusta
TRM	Test Robusto de las Mediciones

Capítulo 7

Conclusiones y Trabajos Futuros

7 Conclusiones y Trabajos Futuros

7.1 Conclusiones

En este trabajo de tesis se han analizado las metodologías existentes para la reconciliación de datos en línea utilizando herramientas de la Estadística Robusta. En este marco se realizó una revisión bibliográfica extensiva que permitió visualizar las limitaciones de las metodologías existentes, así como las ventajas y áreas por explorar.

Es sabido que la Reconciliación de Datos clásica (RDC) permite obtener estimaciones consistentes de las variables de un proceso por medio de la minimización de las discrepancias existentes entre las mediciones y las ecuaciones que lo describen. Está técnica supone que las mediciones se ajustan exactamente a una distribución de probabilidad dada, que por lo general se considera es la normal. No obstante, la presencia de mediciones con errores sistemáticos infringe dicha suposición generando el sesgamiento de los resultados de la RDC. Para hacer frente a esta situación se han desarrollado numerosas estrategias que utilizan test clásicos en conjunto con procedimientos iterativos que buscan detectar, identificar y ejercer acciones correctivas sobre las mediciones clasificadas como atípicas. Estos procedimientos demandan altos requerimientos de cómputo que evitan su utilización en línea.

En la década del 90, la aplicación de la Estadística Robusta a la Ingeniería de Procesos dio origen a la Reconciliación de Datos Robustas (RDR), la cual ha sido aplicada con éxito a mediciones que presentan Errores Sistemáticos Esporádicos (ESE). En este sentido, en la presente tesis se han desarrollados dos estrategias que se denominan Metodología Simple (MSi) y Metodología Sofisticada (MSo) con el fin de obtener estimaciones insesgadas en tiempos cortos. Estas aprovechan las principales ventajas de

los M-estimadores monótonos y redescendientes y hacen uso de la Redundancia Temporal (RT) existente en un conjunto de observaciones. La comparación de MSi y MSo con estrategias presentadas por otros investigadores en la última década permitió concluir que la metodología MSi es una alternativa eficiente, ya que proporcionó buenas estimaciones para las mediciones reconciliadas, y su carga computacional es la más baja gracias a los beneficios de una correcta inicialización, la cual fue obtenida por medio del cálculo de la mediana robusta.

En esta tesis se han desarrollado dos tests para la detección de ESE, denominados Test de las Mediciones de la Ventana (TMV) y Test Robusto de las Mediciones (TRM). Éstos deben su nombre al clásico Test de las Mediciones el cual es capaz de detectar e identificar mediciones con ESE en variables redundantes, aunque lo hace con un elevado número de falsas alarmas. El TMR y el TMV utilizan la redundancia temporal provista por un conjunto de observaciones; gracias a esto consiguen detectar e identificar mediciones atípicas en variables con redundancia espacial nula, con un porcentaje de aciertos idéntico al de las variables medidas redundantes. Esto es un notable avance en las técnicas de Detección de ESE pues independiza la capacidad de detección de la redundancia espacial (RE).

La comparación de las medidas de desempeño del TMV y TRM muestra el deterioro de la exactitud de las estimaciones utilizadas por el TMV, sin embargo este consigue un mayor porcentaje de detección de ESE, pero a expensas de que el número de falsas alarmas aumente significativamente. Esta situación no se produce con el TRM porque emplea estimaciones insesgadas de las variables medidas y su desempeño mejora con el tamaño de la ventana.

La RE de los sistemas lineales puede ser calculada de forma analítica siguiendo desarrollos como el de Maronna y Arcas (2009). No obstante, hasta la fecha, la RE de las variables involucradas en sistemas no lineales, solo podía ser calculada luego de resolver el problema de RD para un conjunto de mediciones. En esta tesis, se extiende el desarrollo propuesto por Maronna y Arcas para calcular la RE en sistemas no lineales. Con estos valores de redundancia se comprobó que el TRM presenta desempeños similares a los obtenidos en sistemas lineales independientemente de la RE de las variables involucradas.

Además, el TRM permite identificar las variables con ESE en sistemas complejos, como procesos con corriente paralelas o variables equivalentes. En los mismos se logran aislar variables problemáticas sin generar falsas alarmas o perder capacidad de detección. Con lo cual se aborda un problema cuya solución estaba pendiente hasta el momento.

El efecto de la presencia de ESE puede ser contrarrestado por la RDR; no obstante, si los Errores Sistemáticos Persisten en el Tiempo (ESPT), las estimaciones se ven afectadas. En este sentido se presenta una nueva metodología para la detección y clasificación de ESPT. Su adecuada utilización en conjunto con la RDR mejora significativamente la exactitud de las estimaciones de las variables. Dicha metodología realiza la RDR con MSi, luego aplica el TRM de forma consecutiva, para detectar mediciones atípicas, y finalmente formula la regresión robusta para clasificar el ESPT. Además, la metodología propone acciones diferentes correctivas para los distintos ESPT presentes. Con esto, se consigue detectar un elevado porcentaje de ESPT y reducir el efecto de los mismos en la RDR. A diferencia de trabajos existentes en este área, se propuso un procedimiento sistemático aplicables a diversos procesos sin modificaciones. Se destaca el análisis exhaustivo de desempeño que se realizó, porque a diferencia de los

trabajos presentados por otros autores, los errores sistemáticos fueron generados de forma aleatoria en todas las variables.

Las estrategias propuestas en esta tesis han sido probadas satisfactoriamente en un modelo de mayor escala como es el de la planta de biodiesel. Por esto, se concluye que la correcta aplicación de la Estadística Robusta al procesamiento de datos permite desarrollar estrategias que proveen de estimaciones insesgadas de las variables, con resultados reproducibles y aplicables a otros sistemas.

7.2 Trabajos futuros

Los posibles trabajos futuros están orientados a ampliar los alcances de esta tesis. A continuación, se citan y explican los mismos.

- Desarrollo de metodologías robustas para detección de errores sistemáticos en las mediciones y pérdidas en el proceso.

La estrategia desarrolla fue probada de forma exitosa cuando los errores están presentes en las mediciones, por el contrario, no se consideró la presencia de pérdidas en el proceso. Es por esto que se planifica ampliar el alcance de la metodología propuesta de forma de detectar, identificar y estimar la magnitud de las pérdidas.

- Desarrollo de técnicas de detección y clasificación para sistemas dinámicos.

La reconciliación de datos ha sido abordada con el fin de aplicarse a optimización en línea; sin embargo, los sistemas dinámicos pueden presentar los mismos inconvenientes que los que operan en estado estacionario, agravados por las variaciones del mismo con el tiempo. Es por esto que se planea abordar el problema de RDR en sistemas dinámicos.

- Desarrollo de metodologías de RDR para modelos con incertidumbre.

La RD considera que los modelos son exactos. No obstante, en los procesos reales existen situaciones de incertidumbre bajo las cuales la RDR no ha sido aplicada. La RDC ha sido estudiada por otros autores en modelos lineales con incertidumbre cuando las mediciones contienen errores aleatorios. Se prevé ampliar este análisis para considerar la presencia de errores sistemáticos. Para disminuir el efecto de estos errores se trabajará con RDR. Además se actualizará el grado de incertidumbre de los parámetros del modelo, pues estos valores podrán ir variando con la evolución del sistema y necesitarán ser restablecidos. Por esto se formulará un problema de RDR con el cual se buscará obtener estimaciones óptimas de las variables de proceso y valores actualizados de los parámetros que se adapten a las nuevas condiciones del sistema.

- Seleccionar un estadístico multivariado que permita detectar fallas.

Se prevé realizar una evaluación rigurosa de las técnicas de identificación multivariadas para detección de fallas. Al utilizar RDR para obtener las estimaciones de las variables, antes de aplicar cualquiera de estos test se deberá verificar que los residuos sigan los supuestos de cada uno de ellos.

- Trabajar con metodologías híbridas para la detección de situaciones anómalas en modelos no lineales con incertidumbre.

Se prevé integrar la información provista por modelos basados en principios de conservación y modelos basados en datos, con esto se busca detectar y diagnosticar las eventos anómalos, respectivamente. Por un lado se aplicara RDR a modelos con incertidumbre lo que permitirá obtener vectores estimados precisos aún cuando el vector de medición contenga valores atípicos. El cálculo de los residuos y las metodologías de control multivariado permitirá detectar las variables sospechosas, las que serán informadas a la etapa

de diagnóstico. Por otro lado, se utilizaran metodologías como Redes Bayesianas y Metodologías Iterativas para diagnosticar la causa real del problema con el fin de poder ejercer acciones tendientes a la recuperación las condiciones operativas normales.

7.3 Acrónimos

ESE	Error Sistemático Esporádico
ESPT	Error Sistemático que Persiste en el Tiempo
RE	Redundancia Espacial
RDC	Reconciliación de Datos Clásica
RDC	Reconciliación de Datos Clásica
RDR	Reconciliación de Datos Robusta
RDR	Reconciliación de Datos Robusta
MSi	Método Simple
MSo	Método Sofisticado
TRM	Test Robusto de las Mediciones
TMV	Test de la Ventana de Mediciones

Referencias

Albuquerque, J. S.; Biegler, L.T. Data reconciliation and gross error detection for dynamic Systems. *AIChE J*. **1996**, 42, 2841-2856.

Almasy, G. A.; Sztano, T. Checking and correction of measurements on the basis of linear system model. *Problems of Control and Information Theory* **1975**, 4, 57-69.

Amand, Th.; Heyen, G.; Kalitventzeff, B. Plant monitoring and fault Detection – Synergy between Data Reconciliation and Principal Component Analysis. *Comput. Chem. Eng.* **2001**, 25, 501-507.

Arora, N.; Biegler, L.T. Redescending estimators for data reconciliation and parameter estimation. *Comput. Chem. Eng*. **2001**, 25, 1585-1599.

Al-Widyan, M. I.; Al-Shyoukh A. O. Experimental evaluation of the transesterification of waste palm oil into biodiesel, *Bio. Tech.* **2002**, 85(3), 253-256

Bagajewicz, M.; Jiang, Q. Gross error modeling and detection in plant linear dynamic reconciliation. *Comput. Chem. Eng*. **1998**, 24, 1789-1809.

Bagajewicz, M. *Smart Process Plants: Software and Hardware Solutions for Accurate Data and Profitable Operations*: *Data Reconciliation, Gross Error Detection, and Instrumentation Upgrade*. McGraw-Hill: New York, 2010.

Britt, H. I.; Luecke, R. H. The Estimation of Parameters in Nonlinear Implicit Models. *Technometrics* **1973**, 15, 233 - 247.

Canakci, M.; Van Gerpen, J. Biodiesel production from oils and fats with high free fatty acids. *Trans. Am. Soc. Agric. Eng.* **2001**, 44 (6), 1429–1436.

Canavos, G. *Probabilidad y Estadística. Aplicaciones y Métodos*. McGraw-Hill: Méjico, 1988.

Charpentier, V.; Chang, L.; Schwenzer, G.; Bardin, M. An Online Data Reconciliation System for Crude and Vacuum Units. Proceeding of NPRA Computer Conference, Houston, 1991.

Chen, J.; Peng, Y.; Muñoz, J. Correntropy Estimator for Data Reconciliation. *Chem. Eng. Sci*. **2013**, 104, 10019-10027.

Cotabarren, N. S. *Ingeniería del Equilibrio entre Fases en Biorrefinerías de Base Oleaginosa*, tesis, Departamento de Ingeniería Química, Universidad Nacional del Sur, 2017.

Crowe, C. M.; García Campos, Y. A.; Hrymak, A. Reconciliation of Process Flow Rates by Matrix Projection Part I: Linear Case. *AIChE J*. **1983**, 29, 881-888.

Crowe, C. M. Reconciliation of Process Flow Rates by Matrix Projection Part II: The Nonlinear Case. *AIChE J*. **1986**, 32, 616-623.

Crowe, C. M. Recursive Identification of Gross Errors in Linear Data Reconciliation. *AIChE J*. **1988**, 34, 541-550.

Crowe, C. M. Test of maximum power for detection of gross errors in process constraints. *AIChE J*. **1989a**, 35, 869-872.

Crowe, C. M. Observability and Redundancy of Process Data for Steady State Reconciliation. *Chem. Eng. Sci.* **1989b**, 44, 2909-2917.

Crowe, C. M. Data Reconciliation - Progress and Challenges". *J. Process Contr*. **1996**, 6, 89-98.

Dennis J.E.; Welsch R.E. Techniques for Nonlinear Least Squares and Robust Regression. *Proc. Am. Statist. Assoc.* **1976**, 83-87.

Drapper, N.; Smith, H. *Applied Regression Analysis*; John Wiley & Sons, Inc.: New York, 1968.

Edgar, T.; Himmelblau, D. *Optimization of Chemical Processes*; Mc. Graw-Hill, 1989.

Freedman, B.; Pryde, E.H.; Mounts, T.L. Variables affecting the yields of fatty esters from transesterified vegetable oils. *J. Am. Oil Soc. Chem*. **1984**, 61, 1638–1643.

Hegel, P.; Andreatta, A; Pereda, S.; Bottini, S.; Brignole, E. A. High pressure phase equilibria of supercritical alcohols with triglycerides, fatty esters and cosolvents, *Fluid Phase Equilib*. **2008**, 266 (1–2), 31–37.

Huber, P. J. Robust Estimation of a Location Parameter. *Annals of Mathematical Statistics* **1964**, 35, 73-101.

Huber, P. J. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematics and Statistics Probability* **1967**, 1, 221–233.

Huber, P. J. *Robust Statistics*; John Wiley: New York, 1981.

Hampel, F. R. *Contributions to the Theory of Robust Estimation*. Ph.D. Thesis, University of California, Berkeley, United States, 1968.

Hampel, F. R. A General Qualitative Definition of Robustness. *Annals of Mathematical Statistics* **1971**, 42, 1887-1896.

Hampel, F. R. The influence Curve and its Role in Robust Estimation. *Journal of American Statistical Association* **1974**, 69, 383-393.

Jacob, J.; Paris, J. Data Sampling and Reconciliation, Application to Pulp and Paper Mills – Part I: Methodology and Implementation", *Appita Journal* **2003**, 56, 25-29.

Jiang, Q.; Sánchez, M.; Bagajewicz, M. On the Performance of Principal Component Analysis in Multiple Gross Error Identification. *Ind. Eng. Chem. Res* **1999**, 38, 2005-2012.

Jiang, Y.; Liu, A. Gross Error Detection and Data Reconciliation Based on A GLR-NT Combined Method. *Journal of East China University of Science and Technology*, 2011.

Johnston, L. P. M.; Kramer, M. A. Maximum Likelihood Data Rectification: Steady-State Systems. *AIChE J* **1995**, 41, 2415-2426.

Jordache, C.; Mah, R.; Tamhane, A.; Performance Studies of the Measurement Test for Detection of Gross Errors in Process Data. *AIChE J*. **1985,** 31, 1187-1201.

Keller, J.Y.; Darouach, M.; Krzakala, G. Fault Detection of Multiple Biases or Process Leaks in Linear Steady State Systems. *Comput. Chem. Eng*. **1994**, 18, 1001-1004.

Kenneth J. Biofuels Annual-Argentina. *USDA Foreign Agricultural Service*, 2017.

Kim, I.; Kano, M.; Park, S.; Edgar, T. Robust Data Reconciliation and Gross Error Detection: The Modified MIMT Using NLP. *Comput. Chem. Eng*. **1997**, 21, 775-782.

Kuehn, D. R.; Davidson, H. Computer Control. II. Mathematics of Control. *Chem. Eng. Progress* **1961**, 57, 44-47.

Kulkarni, M. G.; Dalai, A. K. Waste cooking oil - an economic source for biodiesel: a review. *Ind. Eng. Chem. Res.* **2006**, 45, 2901–2913.

Lam, M. K.; Lee, K. T.; Mohamed, A. R. Homogeneous, heterogeneous and enzymatic catalysis for transesterification of high free fatty acid oil (waste cooking oil) to biodiesel: A review. *Biotechnol. Adv*. **2010**, 28 (4), 500–518.

Lawrence, P. J. Data reconciliation: Getting better information. *Hydrocarbon* **1989**, 68, 55-60.

Liebman. M.; Edgar, T. Data Reconciliation for Nonlinear Processes, *AIChE Annual Meeting*, Washington, D. C., 1988.

Liebman, M.; Edgar, T.; Lasdon, L. Efficient Data Reconciliation and Estimation for Dynamic Processes Using Nonlinear Programming Tecniques. *Comput. Chem. Eng*. **1992**, 16, 963.

Llanos, C.; Sánchez, M.; Maronna, R. Robust Estimators for Data Reconciliation. *Ind. Eng. Chem.* Res. **2015**, 54, 5096-5105.

Llanos, C.; Sánchez, M.; Maronna, R. Classification of Systematic Measurement Errors within the Framework of Robust Data Reconciliation. *Ind. Eng. Chem.* Res. **2017**, 56, 9617-9628.

Lotero, E.; Liu, Y.; Lopez, D. E.; Suwannakarn, K.; Bruce, D. A.; Goodwin, J. G. Synthesis of biodiesel via acid catalysis. *Ind. Eng. Chem. Res*. **2005**, 44(14), 5353–5363.

Ma, F; Clements, L. D.; Hanna, M. A. The effects of catalyst, free fatty acids, and water on Transesterification of Beef Tallow. *Trans. ASAE*. **1998,** 41 (5), 1261– 1264.

McBride, N. *Modeling the production of biodiesel from waste frying oil*. thesis, Department of Chemical Engineering, **1999**, University of Ottawa.

Madron, F. *Process Plant Performance. Measurement and Data Processing for Optimization and Retrofits.* Ellis Horwood Ltd.: Chichester, England, 1992.

Mah, R. S. H.; Stanley, G.; Dowing, D. Reconciliation and Rectification of Process Flow and Inventory Data. *Ind. Engn. Chem. Process Des. Dev.* **1976**, 15, 175-183.

Mah, R. S. H.; Tamhane, A. C. Detection of Gross Errors in Process Data. *AIChE Jl*. **1982**, 28, 828-830.

Mah, R. S. H. *Chemical Process Structures and Information Flows*. Butterworths, 1990.

Maronna, R. A.; Martin, R. D.; Yohai, V. *Robust Statistics: Theory and Methods*; John Wiley and Sons Ltd.: Chichester, 2006.

Maronna, R.A.; Arcas, J. Data reconciliation and gross error detection based on regression. *Comput. Chem. Eng*. **2009**, 33, 65-71.

Martínez Prata, D.; Pinto, J. C.; Lima, E. L. Comparative Analysis of Robust Estimators on Nonlinear Dynamic Data Reconciliation. *Comp. Aided Chem. Eng.* **2008**, 25: 501-506.

Martínez Prata, D.; Schwaab, M.; Lima, E. L.; Pinto, J. C. Simultaneous Robust Data Reconciliation and Gross Error Detection through Particle Swarm Optimization for an Industrial Polypropylene Reactor. *Chem. Eng. Sci*. **2010,** 65, 4943-4954.

Matyus, T.; Gleib, A.; Gruber, K.; Bauer, G.; Data reconciliation structure analysis and simulation of waste flows: case study Vienna. *Waste Management and Research* **2003**, 21, 93-109.

Mei C.; Su, H.; CHU, J. An NT-MT Combined Method for Gross Error Detection and Data Reconciliation. *Chinese J. Chem. Eng*. **2006**, 14, 592-596.

Narasimhan, S.; Mah, R. S. H. Generalized Likelihood Ratio Method for Gross Error Identification. *AIChE J.* **1987**, 33, 1514-1521.

Narasimhan, S. Maximum Power Tests for Gross Error Detection Using Likelihood Ratios. *AIChE J.* **1990**, 36, 1589-1591.

Narasimhan, S.; Jordache, C. *Data Reconciliation and Gross Error Detection*; Gulf Publishing Company: Houston, 2000.

Nicholson, B.; López-Negrete, R.; Biegler, L. T. On-line State Estimation of Nonlinear Dynamic Systems with Gross Errors. *Comput. Chem. Eng*. **2014,** 70, 149-159.

Özyurt, D. B.; Pike, R. W. Theory and practice of simultaneous data reconciliation and gross error detection for chemical processes. *Comput. Chem. Eng*. **2004**, 28, 381-402.

Pai, C.; Fisher, D. Application of Broyden's Method to Reconciliation of Nonlinearly Constrained Data, *AIChE J.* **1988**, 34, 873-87

Rafiee, A.; Behrouzshad, F. Data reconciliation with application to a natural gas processing plant. *Journal of Natural Gas Science and Engineering* **2016**, 538-545.

Ramamurthi, Y.; Bequette, B. Data Reconciliation of Systems with Unmeasured Variables Using Nonlinear Programming Techniques. *AIChE Spring National Meeting*, Orlando, Florida, 1990.

Renganathan, T.; Narasimhan, S. A Strategy for Detection of Gross Errors in Nonlinear Processes. *Ind. Eng. Chem.* Res. **1999**, 38, 2391-2399.

Rey, W.J.J. *Introduction to Robust and Quasi-Robust Statistical Methods;* Springer-Verlag Berlin Heidelberg: Berlin, 1983.

Ripmeester, W.E. *Modeling the production of biodiesel oil from waste cooking oil*. B.A.Sc. thesis, Department of Chemical Engineering, University of Ottawa. 1998.

Ripps, D. L. Adjustment of Experimental Data. *Chemical Engineering Progress Symposium Series* **1965**, 61, 8-13.

Rollins, D.; Davis, J. Unbiased Estimation of Gross Errors in Process Measurements. *AIChE J.* **1992**, 38, 563-572.

Romagnoli J. On Data Reconciliation: Constraints Processing and Treatment of Bias. *Chem. Eng. Sci*. **1983**, 38, 1107-1117.

Romagnoli J.; Stephanopoulos, G. Rectification of Process Measurement Data in the Presence of Gross Errors. *Chem. Eng. Sci*. **1981**, 36, 1849-1863.

Romagnoli, J.; Sánchez, M. *Data Processing and Reconciliation for Chemical Process Operations*; Academic Press: San Diego, 2000.

Rosemberg, J.; Mah, R. S. H.; Iordache, C. Evaluation of Schemes for Detecting and Identifying Gross Errors in Process Data. *Ind. Eng. Chem.* Res. **1987**, 26, 555-564.

Sagar, Y. V.; Tiwari, A. P.; Degweker, S. B. An Iterative Principal Component Test for Fault Detection and Isolation. *Proceedings of 2015 IEEE Multi-Conference on Systems and Control*, Sydney, Australia, 2015.

Sánchez, M. *Monitoreo de Procesos: Análisis de Instrumentación y Reconciliación de Datos de Planta*. Tesis Doctoral Universidad Nacional del Sur, Bahía Blanca, Argentina, 1996.

Sánchez, M.; Romagnoli, J. Use of orthogonal transformations in Data Classification – Reconciliation. *Comput. Chem. Eng*. **1996**, 20, 483- 493.

Sánchez, M.; Romagnoli, J. A.; Jiang, Q.; Bagajewicz, M. Simultaneous Estimation of Biases and Leaks In Process Plants. *Comput. Chem. Eng*. **1999**, 23, 841-857.

Sánchez, M.; Maronna, R. Simple Approaches for Robust Data Reconciliation. 2009 *AIChE Annual Meeting, Nashville, TN*. United States, **2009** Code 79788.

Serth, R.; Heenan, W. Gross Error Detection and Data Reconciliation in Steam Metering Systems. *AIChE J*. **1986**, 32, 733-742.

Singh, S. R.; Mittal, N. K.; Sen, P. K. A Novel Data Reconciliation and Gross Error Detection tool for the Mineral Processing Industry. *Minerals Engineering* **2001**, 14, 809-814.

Sunde, S.; Berg, O. Data Reconciliation and fault Detection by Means of Plant Wide Mass and Energy Balances. *Progress in Nuclear Energy* **2003**, 43, 97-104.

Sun, S.; Dao, H.; Gong, Y. A MT-NT-MILP Combined Method for Gross Error Detection and Data Reconciliation. *IEEE*, **2010**.

Swartz C. L. E. Data Reconciliation for Generalized Flowsheet Applications. *American Chemical Society of National Meeting.* Dallas, TX , 1989.

Tamhane, A. C. A Note on the use of Residuals for Detecting an Outlier in Linear Regression. *Biometrika* **1982**, 69, 488-489.

Tjoa, I. B.; Biegler, L. T. Simultaneous Strategies for Data Reconciliation and Gross Error Detection of Nonlinear Systems. *Comput. Chem. Eng.* **1991**, 15, 679-690.

Tong, H.; Crowe, C. Detection of Gross Errors in Data Reconciliation by Principal Component Analysis. *AIChE J*. **1995**, 41, 1712- 1722.

Tukey, J. W. A Survey of Sampling from Contaminated Distributions. *Contributions to Probability and Statistics*; OLKIN, I.: Stanford University Press, California, **1960**, 448-485.

Tukey, J. W. The Future of Data Analysis. *Annals of Mathematics Statistics* **1962**, 33, 1-67.

Van ser Heijden, R. T. J. M.; Romein, B.; Heijnen, J. J., Linear Constraint Relation in Biochemical Reaction Systems: I. Classification of the Calculability and the Balanceability of Conversion Rates. *Biotechnology and Bioengineering* **1993a**, 43, 3-10.

Van Der Heijden, R. T. J. M.; Romein, B.; Heijnen, J. J., Linear Constraint Relation in Biochemical Reaction Systems: II. Diagnosis and Estimation of Gross Error. *Biotechnology and Bioengineering* **1993b**, 43, 11-20.

Vyas, A.P; Verma, J. L.; Subrahmanyam, N. A review on FAME production processes. *Fuel*. **2010**, 89 (1), 1–9.

Wang, H.; Song, Z.; Wang, H. Statistical process monitoring using improved PCA with optimized sensor locations. J Process Contr. **2002**, 735–744.

Wang, D.; Romagnoli, J. A.; A framework for robust data reconciliation based on a generalized objective function. *Ind. Eng. Chem.* Res. **2003**, 42, 3075-3084.

Wang, F.; Jia, X.; Zheng, D.; Yue, J. An improved MT-NT method for gross error detection and data reconciliation. *Comput. Chem. Eng*. **2004**, 28, 2189-2192.

Yang, Y.; TEN, R.; Jao, L. A Study of Gross Error Detection and Data Reconciliation in Process Industries. *Comput. Chem. Eng*. **1995**, 19, 217-222.

Zhang Z.; Zhijiang S.; Chen X.; Wang K.; Qian J. Quasi-weighted least squares estimator for data reconciliation. *Comput. Chem. Eng*. **2010,** 34, 154-162.

Zhang, Z.; Chen, J. Correntropy Based Data Reconciliation and Gross Error Detection and Identification for Nonlinear Dynamic Processes. *Comput. Chem. Eng*. **2015,** 75, 120-134.

Zhang, Y; Dubé, M.A.; McLean, D.D.; Kates, M.. Biodiesel production from waste cooking oil: 1. Process design and technological assessment. *Bio. Tech*. **2003**, 89, 1–16.

Zheng, S; Kates, M; Dubé, M.A.; McLeana, D.D. Acid-catalyzed production of biodiesel from waste frying oil. *Biomass and Bioenergy*. **2006**, 30(3), 267-272

Zhou, L.; Fu, Y. Data reconciliation based on robust estimator and MT-NT method. *Proceedings of the 35th Chinese Control Conference*, Chengdu, China, 2016.

Apéndice 1

Capítulo 3

Distribución de un M-estimador

La aproximación de la estimación de un M-estimador en muestras de tamaños finito se realiza de forma intuitiva.

Se tiene que si ψ (función influencia) es creciente; para una distribución dada F, se define $\mu_0 = \mu_0(F)$ como la solución de:

$$E_F \psi(x - \mu_0) = 0\,, \tag{A3.1}$$

En general F es simétrica entonces μ_0 coincide con el centro de simetría. Se puede demostrar por la ley de los grandes números que cuando $n \to \infty$

$$\hat{\mu} \to_p \mu_0 \tag{A3.2}$$

donde $\to_p$ significa que tiende en probabilidad a y μ_0 está definida por (A3.1).

La derivación heurística de (A3.2) se puede realizar de la siguiente forma:

Sean las funciones:

$$\lambda(s) = E\psi(x - s); \quad \hat{\lambda}_n(s) = \frac{1}{n}\sum_{i=1}^{n} \psi(x_i - s) \tag{A3.3}$$

donde $\hat{\mu}$ y μ_0 verifican respectivamente que:

$$\lambda(\mu_0) = 0; \quad \hat{\lambda}_n(\hat{\mu}) = 0 \tag{A3.4}$$

Para cada *s*, la variable aleatoria $\psi(x_i - s)$ es independiente e idénticamente distribuidas con media $\lambda(s)$, y por la regla de los grandes números cuando $n \to \infty$ el valor estimado tiende en probabilidad a:

$$\hat{\lambda}_n(s) \to_p \lambda(s) \; \forall \; s \tag{A3.5}$$

Varianza del M-estimador

Partiendo de la expresión:

$$\hat{\mu} = \min \sum_{i=1}^{n} -\ln f_0(\varepsilon_i) = \min \sum_{i=1}^{n} -\rho(\varepsilon_i) \tag{A3.6}$$

donde $\hat{\varepsilon}_i = x_i - \hat{\mu}; \quad i = 1,...,n$ .Si ρ es diferenciable entonces la derivada respecto de μ

$$\frac{d\left[\sum_{i=1}^{n} -\rho(\varepsilon_i)\right]}{d\mu} = \sum_{i=1}^{n} -\psi(\varepsilon_i) = 0 \tag{A3.7}$$

Aplicando la expansión de Taylor a la sumatoria en torno a un $\varepsilon_0 = x_i - \mu_0; \quad i = 1,...,n$

$$\sum_{i=1}^{n} -\psi(\hat{\varepsilon}_i) = \sum_{i=1}^{n} -\psi(\varepsilon_{i0}) + \left.\frac{d\left[\sum_{i=1}^{n} -\rho(\varepsilon_i)\right]}{d\varepsilon_i}\right|_{(\varepsilon_i = \varepsilon_{i0})} \left(\varepsilon_i - \varepsilon_{i0}\right) + O(\varepsilon_i - \varepsilon_{i0}) = 0 \tag{A3.8}$$

Analizando el segundo y tercer término se tiene:

$$\left.\frac{d\left[\sum_{i=1}^{n} -\rho(\varepsilon_i)\right]}{d\varepsilon_i}\right|_{(\varepsilon_i = \varepsilon_{i0})} = \sum_{i=1}^{n} \psi'(\varepsilon_{i0}) \tag{A3.9}$$

$$\lim_{(\hat{\varepsilon}_i - \varepsilon_{i0}) \to 0} \frac{O(\varepsilon_i - \varepsilon_{i0})}{\varepsilon_i - \varepsilon_{i0}} = 0 \tag{A3.10}$$

Por lo que la Ec.A3.8 puede ser reescrita como:

$$0 = \sum_{i=1}^{n} \psi(\varepsilon_{i0}) + (\hat{\varepsilon}_i - \varepsilon_{i0}) \sum_{i=1}^{n} \psi'(\varepsilon_{i0}) \tag{A3.11}$$

Reemplazando $\hat{\varepsilon}_i$ y ε_0 se tiene:

$$0 = \sum_{i=1}^{n} \psi(x_i - \mu_0) + (x_i - \hat{\mu} - (x_i - \mu_0)) \sum_{i=1}^{n} \psi'(x_i - \mu_0) \tag{A3.12}$$

De la cual reordenando se obtiene la expresión:

$$(\hat{\mu} - \mu_0) = \frac{\sum_{i=1}^{n} \psi(x_i - \mu_0)}{\sum_{i=1}^{n} \psi'(x_i - \mu_0)} \tag{A(3.13)}$$

Reescribiendo esta expresión

$$(\hat{\mu} - \mu_0) = \frac{\sum_{i=1}^{n} \psi(x_i - \mu_0)}{\sum_{i=1}^{n} \psi'(x_i - \mu_0)} \frac{n}{n} = \frac{ave(\psi(x_i - \mu_0))}{ave(\psi'(x_i - \mu_0))} \tag{A3.14}$$

$$\sqrt{n}(\hat{\mu} - \mu_0) = \sqrt{n} \frac{ave(\psi(x_i - \mu_0))}{ave(\psi'(x_i - \mu_0))} = \frac{A_n}{B_n} \tag{A3.15}$$

Si analizamos el numerador se tiene:

$$E\left[ave\left(\psi(x_i - \mu)\right)\right] = nE\left[\frac{\sum_{i=1}^{n} \psi(\varepsilon_0)}{n} \right] = n\frac{0}{n} \tag{A3.16}$$

$$Var\left[\frac{\sum_{i=1}^{n} \psi(x_i - \mu_0)}{n} \right] = \frac{1}{n^2} E\left[\sum_{i=1}^{n} \psi(\varepsilon_0) - E\left[\sum_{i=1}^{n} \psi(\varepsilon_0) \right] \right]^2$$

$$Var\left[\frac{\sum_{i=1}^{n} \psi(x_i - \mu_0)}{n} \right] = E\left[\psi(\varepsilon_0)\right]^2 \tag{A3.17}$$

Por lo que la variable aleatoria $ave\left[\psi(\varepsilon_0)\right] \sim \mathcal{N}\left(0, E\left[\psi(\varepsilon_0)\right]^2\right)$

Se analiza, B_n, el denominador de la ecuación (A3.15). Se aplica la desigualdad de Chebyshev y la ley de los grandes números.

La desigualdad de Chevyshev dice:

Sea x una variable aleatoria cuya esperanza es E(x), c un número cualquiera real cualquiera y supongamos que $E(x-c)^2$ existe y es finita. Entonces:

$$P\left[\left|x-c\right| \geq \varepsilon\right] \leq \frac{1}{\varepsilon^2} E(x-c)^2$$

Complementos

$$P\left[\left|x-c\right| < \varepsilon\right] \geq 1 - \frac{Var(x)}{\varepsilon^2}$$

Llevando el complemento a la variable $\psi'(\varepsilon_0)$ y aplicando la ley de los grandes números

$$P\left[\left|\psi'(\varepsilon_0) - E\psi'(\varepsilon_0)\right| < \varepsilon\right] \geq 1 - \frac{E\psi'(\varepsilon_0)(1 - E\psi'(\varepsilon_0)}{n\varepsilon^2} \tag{A3.18}$$

$$\lim P\left[\left|\psi'(\varepsilon_0) - E\psi'(\varepsilon_0)\right| < \varepsilon\right] = 1 \tag{A3.19}$$

Ésto también es válido para el promedio y se puede decir que

$$\begin{aligned} &\psi'(\varepsilon_{i0}) \rightarrow_p E\psi'(\varepsilon_0) \\ &\frac{1}{n}\sum_{i=1}^{n} \psi'(\varepsilon_{i0}) \rightarrow_p E\psi'(\varepsilon_0) \end{aligned} \tag{A3.20}$$

Teorema de Slustky

Sea x_i e y_i , i=1…n, dos va si x $x_n \rightarrow_d x$ e $y_n \rightarrow_p c, c \in \mathbb{R}$

$x_n + y_n \rightarrow_d x + c$
$x_n . y_n \rightarrow_d xc$
$x_n / y_n \rightarrow_d x / c, c \neq 0$

Utilizando este teorema

$$\frac{ave(\psi(x_i - \mu_0))}{ave(\psi'(x_i - \mu_0))} = \frac{\frac{1}{n}\sum_{i=1}^{n}\psi(x_i - \mu_0)}{E\psi(x_i - \mu_0)} \tag{A3.21}$$

La varianza resulta

$$Var\left[\frac{\frac{1}{n}\sum_{i=1}^{n}\psi(x_i - \mu_0)}{E\psi'(x_i - \mu_0)}\right] = \frac{E\psi(x_i - \mu_0)^2}{n\left[E\psi'(x_i - \mu_0)\right]^2} \tag{A3.22}$$

Con lo cual

$$\hat{\mu} - \mu_0 \rightarrow \mathcal{N}\left(0, \frac{E\psi(x_i - \mu_0)^2}{n\left[E\psi'(x_i - \mu_0)\right]^2}\right) \tag{A3.23}$$

Multiplicando ambos miembros por la raíz de *n*

$$Var(\hat{\mu} - \mu_0) = \frac{E\psi(x_i - \mu_0)^2}{\left[E\psi'(x_i - \mu_0)\right]^2} \tag{A3.24}$$

Apéndice 2

Capítulo 4

Distribución del Test Robusto de las Mediciones

Sea $\mathbf{y_j}$ un vector, contiene las mediciones de *I* variables en el tiempo *j*, el cual sigue una distribución $N(\boldsymbol{\mu}, \Sigma)$. El mismo se incorpora a una ventana de datos que ingresa a la RDR utilizando la metodología MSi

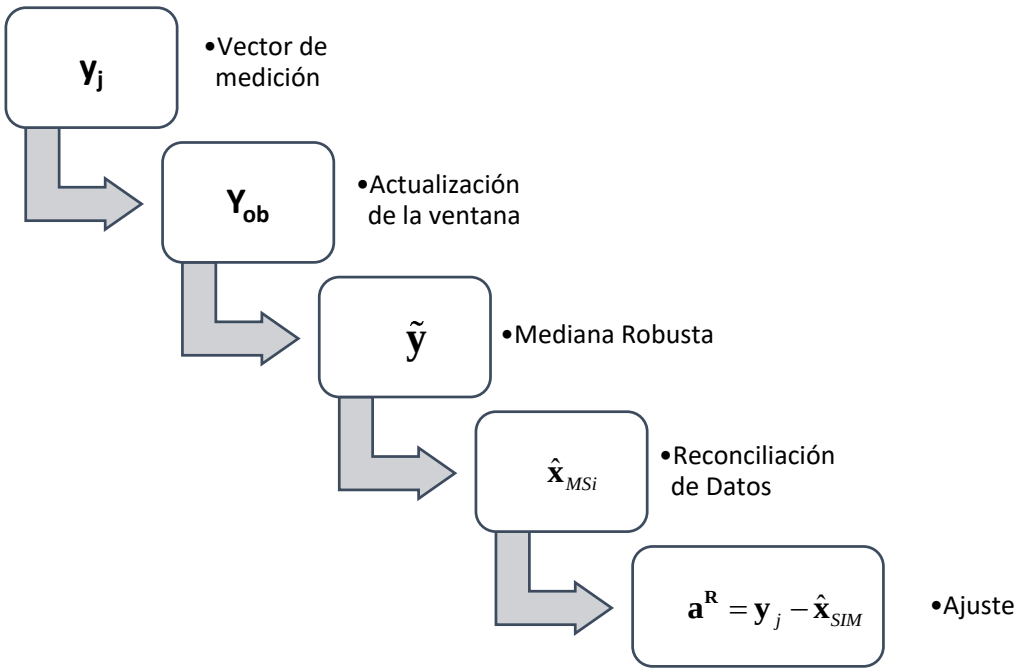


Fig. A4.1. Esquema del cálculo del ajuste robusto

El test de las mediciones se define como la relación entre el ajuste y su varianza. Se sabe que $\mathbf{a^R} \sim \mathcal{N}(0, \hat{\mathbf{Q}}^{\mathbf{R}})$, pero la matriz de covarianza, $\hat{\mathbf{Q}}^{\mathbf{R}}$ es desconocida, por lo que ésta se infiere a partir de una muestra de ajustes.

Con el fin de analizar la distribución que sigue el nuevo estadístico, se estudia para una variable la relación entre el ajuste y la varianza muestral:

$$S^2 = \frac{\sum_{i=1}^{n} (a_i - \mu_a)^2}{n} \tag{A4.1}$$

Como μ_a es desconocida, ésta se reemplaza por la media de los ajustes:

$$S^2 = \frac{\sum_{i=1}^{n}(a_i - \bar{a})^2}{n-1}$$

$$S^2(n-1) = \sum_{i=1}^{n}(a_i - \bar{a})^2 = \sum_{i=1}^{n}(a_{\ i} - \bar{a} + \mu_a - \mu_a)^2$$

$$S^2(n-1) = \sum_{i=1}^{n}\left[(a_{\ i} - \mu_a) - (\bar{a} - \mu_a)\right]^2 \tag{A4.2}$$

Trabajando esta expresión se obtiene

$$S^2(n-1) = \sum_{i=1}^{n}(a_i - \mu_a)^2 - n(\bar{a} - \mu_a)^2 \tag{A4.3}$$

Reordenando y dividiendo en la varianza:

$$\frac{S^2(n-1)}{\sigma_y^2} + \frac{n(\bar{a} - \mu_a)^2}{\sigma_y^2} = \frac{\sum_{i=1}^{n}(a_i - \mu_a)^2}{\sigma_y^2} \tag{A4.4}$$

$$\chi^2$$

Por lo tanto:

$$\Upsilon = \frac{S^2(n-1)}{\sigma_y^2} \sim \chi^2 \tag{A4.5}$$

Asimismo

$$\mathcal{Z} = \frac{a_i - \bar{a}}{\sqrt{Var(a_i - \bar{a})}} = \frac{a_{LS}}{\sigma_{LS}} \sim N(0,1) \tag{A4.6}$$

El test de las mediciones con varianza muestral se calcula como:

$$\tau = \frac{\mathcal{Z}}{\left(\frac{\Upsilon}{df}\right)^{1/2}} = \frac{a_{LS}}{\sigma_{LS}} \frac{1}{\left(\frac{S^2(n-1)}{(n-1)\sigma_y^2}\right)^{1/2}} \tag{A4.7}$$

$$\tau = \frac{a_{LS}}{S} \sim t_{n-1} \tag{A4.8}$$

Luego reemplazando a_{LS} y S por la varianza robusta $\hat{Q}^R$ y el a^R se tiene que:

$$\tau^R = \frac{a^R}{\hat{Q}^R} \sim t_{n-1} \tag{A4.9}$$